KonsortSWD
Konsortium für die
Sozial-, Verhaltens-, Bildungs- und
Wirtschaftswissenschaften

gesis Leibniz-Institut für Sozialwissenschaften

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Task Area 5 Measure 2 Report
Scope: Milestones 1, 2, and 3

Project title	Consortium for the Social, Behavioural, Educational, and Economic Sciences (KonsortSWD)
Task Area 5	Technical solutions
Measure 2	Enhancing data findability
Milestone 1	Study of the searchability and findability of KonsortSWD data on the web
Milestone 1 deliverable	Report
Milestone 2	Development of a strategy to increase the findability of KonsortSWD data on the web (including indicators to measure their impact) based on the requirements derived from the visibility study and using common web standards such as schema.org
Milestone 2 deliverable	Report
Milestone 3	Enrichment of KonsortSWD metadata schemas (e.g., da|ra) with metadata that increase data visibility
Milestone 3 deliverable	Report
Authors	Fidan Limani Yousef Younes Valentina Hiseni Brigitte Mathiak
Reviewed by	Peter Mutschke (incl textual contributions) Janete Saldanha Bach
Date	31 December, 2021

Milestone (1, 2, and 3) Report

Zusammenfassung Deutsch

Die Auffindbarkeit von Forschungsdaten ist ein wichtiger Faktor, um Nachnutzung zu ermöglichen und den Forschungsdatenzyklus zu unterstützen. In diesem Bericht haben wir die Auffindbarkeit von Forschungsdaten unserer KonsortSWD Partner untersucht. Dazu haben wir eine Umfrage unter den Forschungsdatenzentren durchgeführt und Interviews geführt, Web Traffic und Query logs analysiert. Aus diesen Ergebnissen haben wir eine Auffindbarkeitsstrategie und daraus folgend pragmatische Empfehlungen für die Partner entwickelt, insbesondere auch für die Gestaltung der Metadaten. Mit diesem Bericht legen wir die Grundlage für weitere Aktivitäten innerhalb des Measures, um die Auffindbarkeit und Sichtbarkeit unserer Forschungsdaten nachhaltig zu verbessern.

Abstract English

The Findability of research data is an important factor in enabling re-use and support the research data life cycle. In this report, we analyze the findability of data from KonsortSWD partners: we conducted a survey and interviews with the RDC, analyzed web traffic and query logs. From this analysis, we formulate a findability strategy and hands-on recommendations for partners on how to improve their findability, including recommendation on how to shape their metadata. With this contribution, we lay a foundation for on-going activities to sustainably improve the discoverability and visibility of our research data.

Executive Summary (in German)

Trotz des Bekenntnisses zu den FAIR-Prinzipien von allen KonsortSWD Beteiligten hängt die konkrete Umsetzung von "F"indability auf der Detailebene von der individuellen Implementierung ab. So gibt es innerhalb der beiden Gemeinschaften in KonsortSWD - Forscher und Forschungsdatenzentren (RDC) - unterschiedliche Auffassungen, Praktiken und Strategien in Bezug auf die Such- und Auffindbarkeit von Daten im Web. Bevor wir also Vorschläge zur Verbesserung der Such- und Auffindbarkeit von Daten machen können, müssen diese Praktiken und die daraus entstehenden Anforderungen verstanden werden.

Im Folgenden werden wir der Ergebnisse aus verschiedenen Studien zur Auffindbarkeit der Forschungsdaten von KonsortSWD-FDZ präsentieren. Insbesondere auch Indikatoren zur Messung ihrer Auswirkungen und die Verbreitung gemeinsamer Standards wie Schema.org.

Um zu diesen Ergebnissen zu gelangen, wurden verschiedene Methoden eingesetzt, wie z. B. die Überprüfung des aktuellen Stands der Technik zu diesem Thema, die Durchführung eines quantitativen und qualitativen Anforderungsanalyse und zuletzt die Anwendung von Suchmaschinenoptimierung (über ein Tool), um sowohl die zuvor erwähnten Einzeltechniken (die Umfrage und die Interviews) zu unterstützen als auch die einzelnen Gemeinschaften im Hinblick auf die Art und Weise, wie die Datensätze ausgedrückt werden können, zu vertiefen (der SEO-Prozess).

Unsere Untersuchungen beziehen sich auf 34 FDZ aus den Sozialwissenschaften, die Teil des KonsortSWD-Projekts sind. Dabei im Fokus standen Praktiken der Such- und Auffindbarkeit über alle Disziplinen hinweg, die durch eine Umfrage und 7 Interviews erhoben wurden.

Den Umfrageergebnissen zufolge geben 88% der Befragten an, dass sie ein FDZ, ein Datenrepository oder eine Website für ihre Daten nutzen. In 95% der Fälle können sie sich KonsortSWD als potenzielle Unterstützung für ihre Datenauffindbarkeit vorstellen. Darüber hinaus haben wir erfragt, welche FDZ über SEO-Berichte verfügen, welche Forscher, die mit den FDZ verbunden sind, für ein gezielteres Follow-up (Interview) zur Verfügung stehen. Dadurch wurden auch Kandidaten ermittelt, die daran interessiert sind, mit uns bei der Prototypisierung der Ergebnisse unserer Maßnahme zusammenzuarbeiten, die in weiteren Projektphasen geplant ist.

In den Interviews schienen die Teilnehmer im Allgemeinen mit dem Thema (Such- und Auffindbarkeit) vertraut zu sein, da sie zumeist einige Methoden bereits anwenden und oft auch schon die eine oder andere Herausforderung identifiziert haben. Insbesondere besteht ein gemeinsames Grundverständnis der wichtigen Rolle die Metadatenbeschreibungen, Metadatenstandards und einschlägige Praktiken (z. B. die Verwendung von Sitemaps), damit ihre Nutzer ihre (Meta-)Datensammlungen spielen. SEO-Optimierung und -Tools waren weitere Punkte, die im Feedback ernannt wurden. Viele FDZ werden solche Tools bereits an (z. B. Google Search Console und Matomo), auch wenn es zum Teil noch Bedenken gibt beispielsweise bezüglich der Datenschutzkonformität oder wie diese Tools optimal eingesetzt werden können. Ein wichtiger Punkt dabei ist der Mangel an Ressourcen, insbesondere Personalressourcen, um die die Vorteile von SEO vollständig nutzen können.

Einige Empfehlungen, die wir an die Teilnehmende weitergeben haben, beinhalten die Übernahme von Metadatenstandards und die Bereitstellung hochwertiger Metadatendokumentation, die Anwendung von SEO auf den Websites der FDZ, der Einbindung von Metadaten-Aggregatoren und die Unterbringung der FDZ-Website unter einer institutionellen Domain, um den Nutzerverkehr zum FDZ selbst zu erhöhen. Das vielleicht wertvollste Ergebnis des Interviews war es, dass wir Partner gefunden haben, mit denen wir bei der Thematik Datensuche und der Auffindbarkeit von Daten zusammenarbeiten können, auch in den folgenden Projektphasen. Dies ist insbesondere wichtig für Aufgaben, die auf eine solche Beteiligung angewiesen sind, wie z. B. die Entwicklung von Prototypen, die Überwachung der Community usw., wie in diesem Bericht weiter ausgeführt wird.

Dieser Bericht umfasst alle drei Projektmeilensteine des ersten Projektjahres: [M1] Untersuchung der Durchsuchbarkeit und Auffindbarkeit von KonsortSWD-Daten im Web, [M2] Entwicklung einer Strategie zur Verbesserung der Auffindbarkeit von KonsortSWD-Daten im Web auf der Grundlage der aus der Studie in M1 abgeleiteten Anforderungen und [M3] Anreicherung der KonsortSWD-Metadatenschemata mit Metadaten, die die Sichtbarkeit der Daten erhöhen.

Im ersten Teil werden die Ergebnisse der Umfrage, der Interviews, der SEO-Berichterstattung mit dem SISTRIX-Tool und der Keyword-Analyse vorgestellt, die mit einigen der FDZ-Mitglieder durchgeführt wurden. Im zweiten Teil wird sich auf den Umgang mit den lokalen und globalen SEO-Anforderungen der Community, was bei einem Projekt, an dem FDZ aus vielen Institutionen teilnehmen, besonders wichtig ist, konzentriert und schlagen Aktionspunkte vor, um die Sichtbarkeit der Daten für die KonsortSWD-Community zu verbessern (Entwicklung eines Prototyps, Einrichtung eines Community-weiten Monitorings sowie Verbreitungs- und Marketingaktivitäten). Im dritten Teil schließlich werden praktische Empfehlungen aufgeführt, die von den Verantwortlichen direkt als Teil ihrer Datenauffindbarkeitspraktiken übernommen werden können.

Executive Summary

Despite a general commitment to the FAIR principles, the concrete implementation of "F"indability is on the detail level up to the individual implementation. Within both communities in KonsortSWD - researchers and Research Data Centers (RDC), different understanding, practices, and engagements towards searchability and findability of

data on the Web exist. As such, before suggesting anything to improve data searchability and findability, we needed to understand those practices and requirements. As a follow up, we focused on applying the results from the findability studies of KonsortSWD RDCs - including indicators to measure their impact - and using common standards such as Schema.org.

Reaching this understanding set us on a certain methodology to reach this goal. As part of the methodology, we adopted few techniques, such as review state of the art on the topic, conduct a requirements-gathering process - both quantitative and qualitative, and, as a final step, apply Search Engine Optimization to both support the individual techniques mentioned before (the survey and the interviews), or explore individual communities to a more depth with regards to the way to express the datasets (the SEO process).

Our methodology targeted 34 RDCs from social sciences, part of the KonsortSWD project. In doing so, we followed the current searchability and findability practices, across disciplines, and then, conducting a survey and 7 interviews, focused on those adopted by the KonsortSWD communities.

According to the survey results, 88% of the respondents report that they use an RDC, a data repository, or a website for their data. Moreover, in 95% of the cases, they see KonsortSWD as a potential support to their data findability efforts. Finally, we found out which RDCs have SEO reporting, the researchers associated with the RDCs available for a more focused follow up (interview), and we identified candidates interested to collaborate with us in prototyping the findings from our measure.

When it comes to the interview, the participants seemed generally familiar with the topic (of searchability and findability) based on some of the practices they adopt and some of the challenges they face. Namely, there is a common understanding of the role that metadata description, metadata standards, and relevant practices (the use of sitemaps, for example) play in order for their users to find and understand their (meta)data collections. SEO optimization and tools was another point of feedback; while they were, by and large, already relying on such tools (e.g., Google Search Console and Matomo), planning to, or familiar with, data protection concerns and lack of resources dedicated to such operations was preventing them to fully utilize the benefits that SEO brings.

Some of the recommendations we provided include adopting metadata standards and providing rich metadata descriptions, applying SEO to their RDC's websites, relying on metadata aggregators, locating RDC's website under a more popular, institutional domain to increase user traffic to the RDC itself, to name but some of the broader categories. A final and valuable outcome of the interview was that we were able to identify partners to collaborate with on data search and findability practices as we move forward in this KonsortSWD measure. This is especially important for tasks that rely on such user involvement, such as prototype development, community monitoring, and so on, which we will see in this report.

The scope of the work reflected in this report includes three project milestones: [M1] Study the searchability and findability of KonsortSWD data on the Web, [M2] Develop a strategy to increase the findability of KonsortSWD data on the Web based on the requirements derived from the study in M1, and [M3] Enrichment of KonsortSWD metadata schemas with metadata that increase data visibility.

In the first part, we present the results from the survey, interviews, the SEO reporting using the SISTRIX tool, and the keyword analysis we did with some of the RDC members. In the second part, we focus on managing the local vs global SEO requirements of the community, which is especially important to address in a project where RDCs from many institutions participate, and propose action points to improve the data visibility for the KonsortSWD community (developing a prototype, establishing community-wide monitoring, and dissemination and marketing activities). Finally, in the third part, we provide more hands-on recommendations that the communities can adopt as part of their data findability practices.

1 Milestone 1: Study of the searchability and findability of KonsortSWD data on the web

1.1 Introduction

Data discovery is a complex process, involving many steps and scenarios, as Fig. 1 shows. To find datasets, users typically follow multiple paths, such as based on the information about them as cited in a publication, or relying on search engines, aggregators, portals, and so on, dedicated to datasets. Regardless of the path taken, what remains of importance to research communities - including those from social sciences - are the practices they apply in the context of searchability and findability of their (meta)data collections.

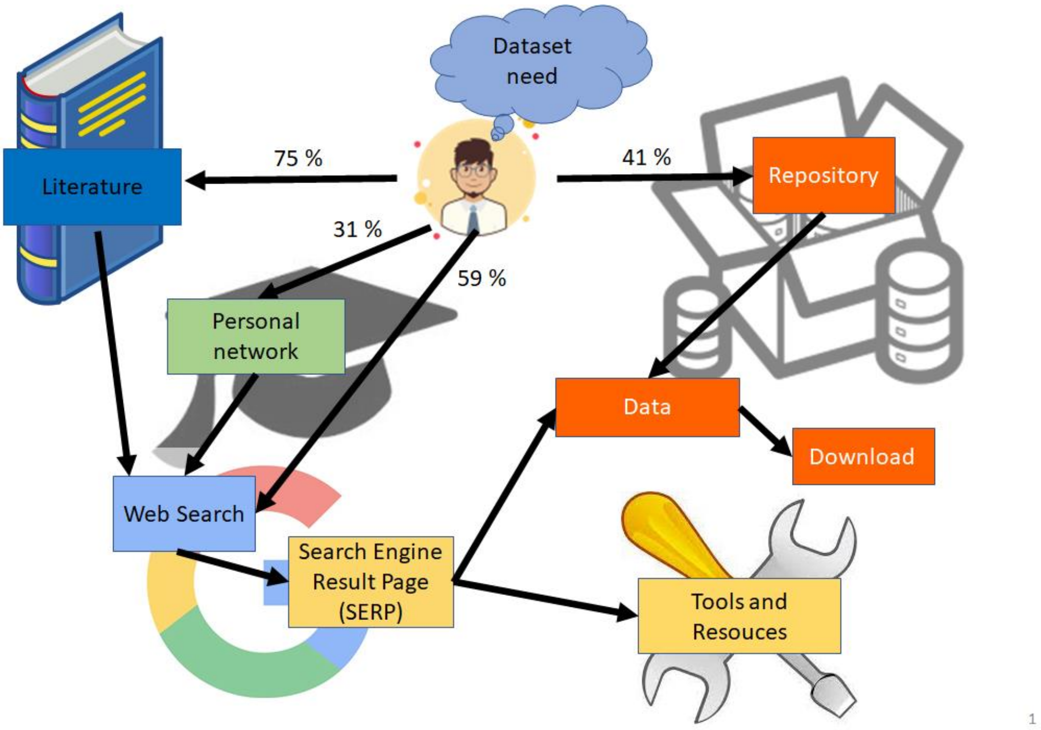


Figure 1 Schematic search path for data discovery: Literature, search engines and data repositories remain the top-3 sources for data to researchers (percentages from Gregory et al., 2020)

When we factor in the specifics of the different research communities, including those for data search, we get another layer of complexity in addition to the one just mentioned about the different paths for data discovery. This is the case even for communities within the same or in relatively close domains, as we will see later in this report. To this end, and focusing on social sciences, for the first project milestone, we conducted a study on the data searchability and findability of the KonsortSWD community, which included 34 Research Data Centers (RDC) that are part of this project. The expected results from this study are understanding the practices around research data that this community uses, identify their current gaps or improvement opportunities, as well as address - or, if not possible, give suggestions to - their current data search challenges.

The methodology we chose to study data searchability and findability involves 4 parts:

- **Survey** Having 34 RDCs to study is a considerable feat. To be able to get as broad of a feedback as possible from their representatives, we selected the survey as our starting point in this process. It represented a way to gather an initial - to some extent structured - feedback, at a certain depth. It was with this feedback that we started to build our understanding and further study directions for this community.

- **Interview** On a more qualitative note, we wanted to get more in-depth information - mainly the rationales for the practices reported via the survey, so we could understand what suggestions or recommendations to give them. There were almost 10 respondents willing to share different aspects around handling research data, and collaborate with us.

- **SEO Analysis** Search Engine Optimization (SEO) is at the core of our goals in the measure. Being that it is common to search for data on the Web, adopting SEO as guiding principles for searchability and findability is an important consideration. As such, we wanted to show the role of SEO to the KonsortSWD community, and potentially work with them on their individual cases with respect to it. We did this by relying on SEO best practices and a tool that we used to analyze the websites of the KonsortSWD RDCs.

- **Keyword** analysis One of the important (metadata) practices for making data searchable and findable includes their rich (metadata) description (also confer F2 of the FAIR principles Wilkinson et al., 2016). Dataset keywords play a significant role in helping users understand a dataset, and for that, we conduct an analysis of keyword usage in describing dataset collection for some of our project members. In addition to the keyword trends, the results of this analysis show the possibility of using logs of user Web searches to derive these trends.

Survey
Interview
Keyword Analysis
SEO Analysis

Figure 2 Approaching data findability practices in KonsortSWD: The methodology

The proposed set of techniques enabled us to engage with the KonsortSWD community, gather their feedback, and use SEO and keyword analysis as motivating material during the requirements gathering process. We will provide more details of each technique in the relevant subsections.

1.2 Search Engine Optimization Insights

Findability in the context of the world wide web refers to the situation in which a web page can be found. There are two types of findability: external and internal (Onaifo et al., 2013). While external findability indicates the situation where users find the webpage by establishing a query against a web search engine such as Google, internal findability refers to the case where the user is able to find a specific webpage from a website while he/she is visiting it. The need for findability originated in the e-commerce field where manufacturers and traders want to promote their products. The first law of e-commerce states that "If the user can't find the product, the user can't buy the product." (Jakob et al., 2001). But soon it spread to other fields, including academia. In academia, research data is the product on which the first law of e-commerce holds with RDCs as the traders and researchers as customers. Although research data are difficult to gather, they are key to reproducibility of experiments and to comparability of results among researchers who together with RDCs agree on one goal that is to make research data findable (Wilkinson et al., 2016). Unfortunately, most of the webpages on the internet are not designed with web findability in mind. That is because developers usually tend to overlook this feature during development either because it is not requested by the user, or because they are not familiar with its requirements. Moreover, speaking about the requirements of discoverability, we should mention that it covers many aspects, such as information architecture, user interface, accessibility and Search Engine Optimization (SEO) (Baker, 2013). This section focuses on the SEO aspect of the discoverability requirements.

The goal of SEO is to improve the findability of a webpage for customers by improving its rank on the Search Engine Result Pages (SERPs). This competition for higher ranking turns out to be a crucial factor for findability. Because users tend to click on the top links that they get as a result of their query, pages that appear later in the list have less chance to be visited by the user. There are many SEO techniques to improve this ranking, such as using important keywords at specific positions of the webpage, using sitemaps, providing links between web pages, etc. In order to decide which technique to use and how to use it on a particular website, we need to analyze that website and compare it to its competitors. This analysis provides the necessary information that guides the optimization process and is usually done using an SEO tool. There are several tools available in the market like Google Search Console, eTracker, SISTRIX, and more. We have chosen SISTRIX as the SEO tool to conduct these analyses.

SISTRIX is an SEO tool that was introduced in 2008. It provides a lot of features and functionalities for SEO professionals to use for analyzing and improving the findability of websites on major web platforms such as Google, Amazon, Tiktok, Instagram and YouTube. For example, it enables the rating of a website and many other competitor sites very easily. It also records/collects/etc., new keywords that can help improve website competitiveness. However, there is one feature that enables measuring SEO success and is only provided by SISTRIX. That feature is the *visibility index*. In addition to this feature, SISTRIX has many more - some of which we explore in this report - that set it as an important tool for companies of different sizes and from different sectors. That is why we choose it as the tool to use for our optimization goals.

In our case, the product which we are interested in improving the findability for, is research data. This research data is provided by KonsortSWD partners and listed in the Survey section. We will use SISTRIX to analyze the websites of a couple of thoses partners. One last thing to mention is that we have set Google as our target search engine because research data is mostly available in English and Google is by far the most widely used search engine.

This SEO report uses SISTRIX to compare the domains of some of the KonsortSWD partners and to show some of their features. In it, we start by comparing their visibility index, which, as a key feature of this tool, takes an important part in this report. Then, we compare the number of keywords and URLs between the domains, after which we show that domains and subdomains have different interesting rankings for different keywords using GESIS and some of its subdomains as an example. We chose GESIS, because there we had the most insight about the structure and additional data on user tracking. However, it turned out that such specialized knowledge wasn't actually necessary to interpret the data. Following, we compare the partners visibility across countries. Finally, we conclude by showing that there is no competition between partners by listing the top keyword competitors, again, using GESIS as an example. This information is a reliable source to understand the strengths and weaknesses of the domain, which is useful for directing the optimization efforts in the right direction.

1.2.1 Web domain visibility

SISTRIX introduced the visibility index as one of its core features. It is a reliable and transparent indicator that allows measuring SEO success on Google. It shows whether the SEO activities were successful or not and what impact had Google updates on the domain's rankings.

The visibility index shows a value for a domain's visibility, for a given date, in the Google organic search results. It can be interpreted as a market-share in the form of an index value that can be compared between domains. The calculation of the visibility index involves three steps. First, collect organic search results for one million keywords (queries) that represent the search volume in the country under focus. Usually 100 results for each keyword, in other words we have 100 M data points. These keywords are not fixed, instead they are changed regularly according to the demand. Second, use the keywords' search volume and the expected click probability on the measured, organic position to weight these keywords.Third, add the weighted values for all rankings of a domain to get the visibility index.

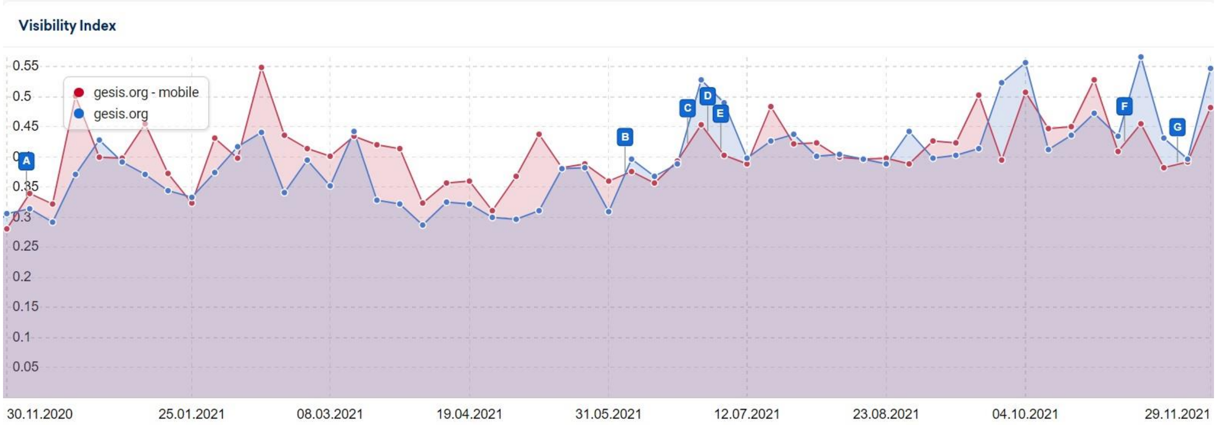


Figure 3 An example visibility index for mobile and desktop

The Google search results for desktop and mobile are different and may depend on a number of things, including mobile usability and the mobile-first indexing algorithm. The underlying keyword set is identical to ensure the comparison between desktop and mobile indexes. Fig. 3 shows an example comparison of mobile and desktop visibility index on GESIS' domain for the last year (showing only one year provided the best trade off between the level of details vs information on the graph). The figure indicates that GESIS' visibility Index on mobile is better than on desktop except at specific dates which correspond to letters. The letters indicate google updates and their impact on both indexes. For example, update c had positive effect on the desktop visibility measure but negative effect on the mobile one.

1.2.2 SEO Metrics: Visibility, keywords and the number of URLs

In the context of SEO, we want to continuously measure the impact that certain optimizations or SEO techniques are having on our websites. As new SEO approaches emerge, or the search engines change the way they rank our websites, we need a way to measure the success of our chosen SEO strategy. SEO metrics enable this by providing insights of interests that help us analyze the organic search results for a given SEO approach (Sickler, 2021). In this section, we will use SISTRIX to collect SEO metrics like visibility, the number of keywords and URLs. These metrics serve as useful tools to compare between domains and their subdomains, and also compare between different domains.

The visibility index is one important factor that can be used for our comparison purposes. It gives a good idea about the visibility of the domains under focus. Table 1 shows these comparisons between some of the KonsortSWD partners in Germany. As an example of the comparison in the same partner, the gesis.org domain has a visibility index of (0.5035) that is much higher than those for search.gesis.org and dbk.gesis.org which are (0.0429, 0.015) respectively. Looking closely at the domains and subdomains of the partners, we can conclude that, in general, the main domains are always more visible than the datasets they host or provide. Of course, this is not a rule because the Social Science Open Access Repository (SSOAR), for example, has a higher visibility index than the main GESIS domain (gesis.org). In addition, comparing the main domains between the different partners listed in Table 1, it can be noticed that IDS has the highest visibility index followed by GESIS.

There are two other factors that can be used to compare and explain the discoverability of different domains; those are the number of keywords and URLs. The number of keywords for which a domain has rankings has an important influence on the visibility of the domain in Google's search result. This number is based on SISTRIX's extended database which contains millions of keywords and it is strongly influenced by the number of URLs with rankings. The number of URLs with rankings refers to the number of URLs that rank for at least one keyword in Google's top 100 results. Table 1 presents these two numbers, along with the visibility index, for the main domain then for the most visible subdomains of some KonsortSWD partners. It shows that the IDS main domain has the largest number of keywords followed by GESIS. But GESIS has the highest number of URLs followed by IDS. But since the visibility index for IDS is higher than that of GESIS, it can be implied that the number of keywords for which a domain has a ranking has more influence than the number of URLs.

Institution	domain/subdomain	Visibility index	Keywords #	URL #
GESIS	gesis.org	0,5035	77.300	16.118

	ssoar.info[1]	1,136	257.100	49.836
	search.gesis.org	0,0429	5.476	4.019
	dbk.gesis.org[2]	0,015	12.291	3.777
ZBW	zbw.eu	0,0855	8.900	4.494
	econbiz.de[3]	1,652	190.183	100.920
	journaldata.zbw.eu	0	69	44
IDS	ids-mannheim.de	1,392	59.341	6.514
	agd.ids-mannheim.de/index.shtml	0,0119	874	75
DIPF	dipf.de	0,3389	20.528	6.085
	fdz-bildung.de	0,0024	2.645	739
DZHW	www.dzhw.eu	0,0324	4.601	1.432
	metadata.fdz.dzhw.eu	0,0003	229	128

Table 1 Visibility Index of domains and subdomains of KonsortSWD Partners. Higher values indicate higher visibility. The domain/subdomain column lists the main domain of the institute and then some important subdomains sorted descendingly by visibility. Keywords # column is the number of keywords for which the domain has a ranking. URL # column is the number of URLs for which the domain has a ranking.

[1] SSOAR.info is the Social Science Open Access Repository

[2] DBK is the former Data Catalog which can be accessed now through GESIS search

[3] econbiz.de is a domain supported by ZBW and provides research related services

1.2.3 Interesting Rankings

In this section, we show that domains and subdomains have different rankings for different keywords by taking GESIS and its subdomains as an example. The keyword lists contain keywords for which either the ranking position is good or the search volume is high. The reason behind choosing these keywords is that they have the potential to bring a large number of visitors to the website via Google search.

Table 2 gives an overview of rankings which are especially interesting for the GESIS main domain and two of its subdomains, those are search.gesis.org and dbk.gesis.org. This is done by choosing keywords from this domain's rankings, which are well positioned and show a large search-traffic and strong competition. Only the top ranking keywords were chosen and listed in a descending order based on the ranking column.

Domain	Keyword	Ranking	Clicks	Search Volume	Competition
gesis.org	cumulation	18	77	15400	51%
	german friendship	6	44	900	27%
	switzerland abbreviation	4	29	400	30%
	socio demographic	3	28	250	27%
	austria abbreviation	3	25	200	21%
search.gesis.org	selbstwirksamkeitserwartung	12	12	1450	40%
	consumer expenditure survey	9	63	2450	32%
	gesis datenbank	7	10	250	37%
	effektive inzidenz	6	7	150	28%
	npi test	4	13	150	26%
dbk.gesis.org	oskar lafontaine ehepartnerin	18	18	3500	43%

	hddsddna st ranka	15	29	4500	44%
	dbk	9	56	2150	61%
	world values survey	8	26	850	24%
	öllieferländer verband	5	18	250	28%

Table 2 Interesting Rankings for some of the GESIS-related domains, sorted descendingly by rankings.

The columns in the table are as follows. Ranking is the position of the domain URL for the keyword inside the organic SERPs. Clicks is the estimated number of organic clicks per month for the keyword. Search Volume is the average number of monthly queries taken over a year for this keyword on Google. Competition indicates, with values between 0 and 100, how difficult it is to rank for the keyword in organic google search.

1.2.3.1 Visibility across countries

The visibility in different countries shows a comparison of the Visibility Indexes of the domain for different Google search countries such as. Google.ie for Ireland, Google.co.uk for the UK and Google.com for the USA. Since the measurement uses a country-specific keyword set for each country, the values are not comparable directly. However, they are useful for checking the functionality of any planned geo-targeting.

Country Partner	DE	AT	Ch	UK	US	BE	IE
gesis.org	0.50	0.51	0.68	0.02	0.01	0.03	0.09
zbw.eu	0.12	0.21	0.29	0	0	0	0
ids-mannheim.de	1,15	1,22	1,27	0	0	0	0
dipf.de	0.31	0.47	0.39	0	0	0	0
dzhw.eu	0.06	0.17	0.10	0	0	0	0

Table 3 Visibility of some KonsortSWD partners across countries

Table 3. shows that KonsortSWD partners are mostly visible from the German speaking countries. The internationality aspect is only available in a weak form for GESIS. One obvious reason for this is the language. This might indicate some issues with the English version of the website that prevents it from being visible in other countries.

1.2.3.2 Top 100-keywords Competitors

Based on the existing rankings of the domain, the keyword competitor list shows other domains that also have rankings for these keywords. The rankings of the examined domain are set equal to 100%. A value of, e.g., 200% means that the associated domain gets about twice as much visibility for the keywords that the surveyed domain rankings in the top 100 have. Accordingly, a value of 50% is only half as much visibility as the examined domain.

Some listed domains are unlikely to be direct competitors of business model, company size, or target audience. However, they are competing on Google for the same keywords and must therefore be considered keyword competitors.

Host	Competition
www.destatis.de	112,1%
www.gesis.org	100%
link.springer.com	79,4%
www.grin.com	56%
www.researchgate.net	44,6%
www.ssoar.info	36,4%
ec.europa.eu	29,8%
www.degruyter.com	29,8%
www.surveymonkey.de	28,8%

Table 4 Top 100-Keyword competitors for gesis.org

Table 4 shows the top 10 competitors for gesis.org for the top 100 keywords. By skimming the list of top 100 competitors, we have encountered no competition with other KonsortSWD partners. But we have noticed a competition with subdomains like ssoar.info, search.gesis.org and dbk.gesis.org which have competition values of 36.4%, 11.6% and 10% respectively. This is another confirmation that the main domain is more visible than the subdomains.

1.2.4 Conclusion

In this section, we have depended on SISTRIX as a reliable tool that enabled us to analyze the domains of KonsortSWD partners from an SEO perspective and compare them. The takeaways conclusions of these comparisons are as follows:

- KonsortSWD partners are not competitors to each other. However, some partners have multiple subdomains, which are competitors to themselves.

- Main domains are generally more visible than their subdomains, but the subdomain inherits visibility nonetheless.

- All the RDC we looked at have a strong focus on German-speaking countries as their primary user group.

- The SISTRIX visibility index corresponds roughly to what we already know from user tracking of the respective domains. It can therefore be used as a tool to track SEO improvements over time.

1.3 Survey

In order to understand the data findability for the KonsortSWD community, we adopted a quantitative approach. This would help us involve as many participants from these communities as possible and share their practices about making their research data findable. We now present the technique adopted to address this quantitative requirement, discuss its initial findings, before concluding the section.

1.3.1 Survey design

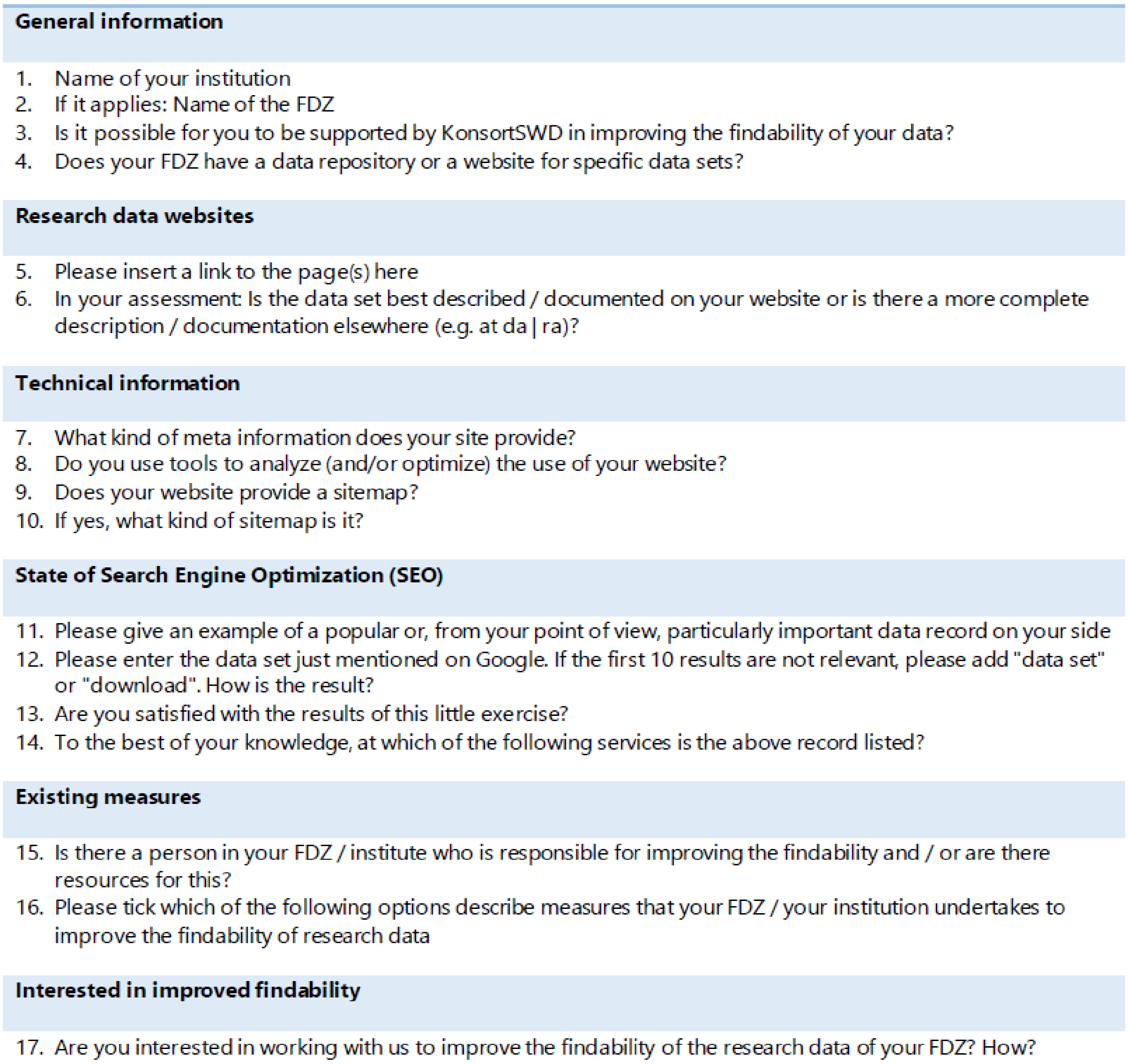
The target group for the survey were KnsortSWD community members with knowledge of data findability in their institution. The sample included members involved in any capacity with relevant tasks on this topic. To this end, we contacted all KonsortSWD members with an explanation of the survey goals. This preparatory phase identified those who were suitable to participate in the survey, such as staff members in charge of or dealing with research (meta)data. Moreover, the invitees were encouraged to share the survey invitation with colleagues that fit the survey target, just in case the invitation did not reach some of them.

Previous work on data search and findability shows that data discovery is a complex process (Krämer et al., 2021; Gregory et al., 2020; Friedrich, 2020), and, to add to this complexity, the specifics of the individual research fields can affect this process. Part of the measure focuses on identifying these aspects for KonsortSWD members.

Table 5 shows the questionnaire. The survey groups the questions according to section themes (highlighted text) to arrange group-related questions. Such an organization provides more information to the participant about the focus of specific questions. On the "administrative" side, the survey ran from April 15 to April 30 2021, and it was sent to 54 participants. The survey response rate was 42% (n = 23), which we assessed as satisfactory as an initial step towards understanding the data findability practices adopted in these communities.

The survey consisted of 17 questions, organized into 6 sections. The first section focused on basic information about the institution, its RDC or website for research data - if one existed, as well as the perceived value from KonsortSWD. With the second section, we were interested in better understanding the data description and documentation for participants' RDCs or websites to identify which metadata standards to consider in the context of data findability. For some communities, there are other services such as RDCs, registries, discipline repositories, and the like - that better describe their data collections. We prompted the participants to point these out in case they existed. Regardless of which entity documents a data collection the best, we were interested in exploring it. Next, we focused on the technical information that communities use to support their data findability, which includes: any metadata standards they rely on, SEO tools (see the next section), any common recommendations, including a sitemap on a website to ease data discoverability. For participants that mentioned SEO as part of their toolbox, we dedicated one section to it.

Section 4 of the survey provided us with details about SEO tools that different communities used, and any practice to support their decision-making about relevant aspects of their dataset publications (metadata descriptions, metadata standards to adopt, keywords to use to describe them, and so on). In this section, through a bit of little task/exercise, the participants were able to assess the discoverability of their data collections. To put these results in perspective, we asked them next (Section 5) about the resources they dedicate to support data findability, and the techniques they use to do this. In the final section of the survey, Section 6, given the current practices, tools, and resources, we wanted to know if the communities were interested in collaborating with us and following our suggestions and findings from the project.



General information

1. Name of your institution
2. If it applies: Name of the FDZ
3. Is it possible for you to be supported by KonsortSWD in improving the findability of your data?
4. Does your FDZ 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 analyze (and/or optimize) 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 FDZ / 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 FDZ / 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 FDZ? How?

Table 5 The data findability survey

1.3.2 Survey results

As we mentioned, the survey contains both "administrative" and data findability aspects, so we will discuss some of the survey results that pertain more to the latter. These aspects include the (a) use of sitemaps, (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. Let's discuss them briefly in order.

(a) Sitemaps play a significant role in data findability, as for any resources on a website, for that matter. The survey informs us that one-third of the participants report using it in the websites or RDCs used to publish their data collections. 42% report not using sitemaps, while the remaining part (26%) are unsure about it (Fig. 8). For some of the participants who reported not using a sitemap for their RDC or website, we found that a sitemap actually exists. This potentially tells us that either they were not aware of a sitemap (26% of the cases), or the solutions they used to publish their collections (such as a Content Management System (CMS), or the like) automatically provide a sitemap about which they are not aware of.

(b) As far as the use of SEO tools goes, we found a variety of such tools, albeit with low usage practice across the participants, as Fig. 9 shows. While 56% of the participants do not yet use an SEO tool, the rest showed 5 different alternative tools (2 - Search Console and Analytics - from Google), which was interesting to observe (and helpful when preparing the interviews).

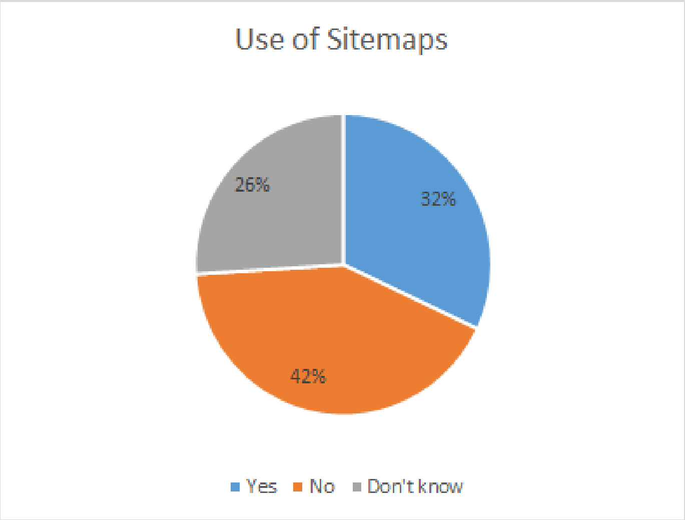


Figure 4 Use of sitemaps

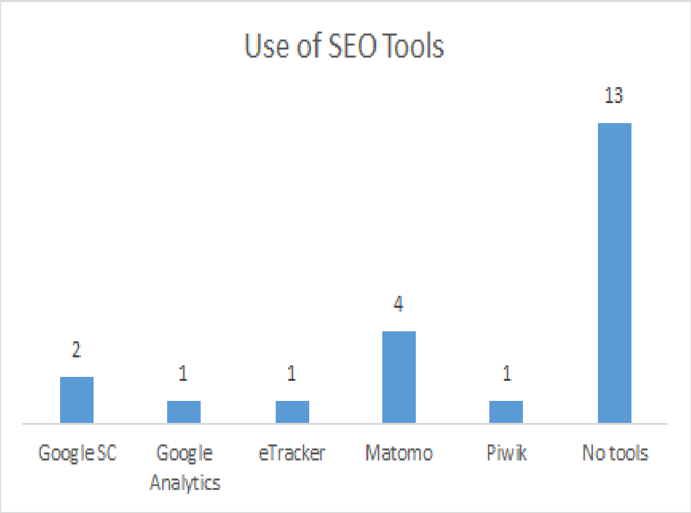


Figure 5 Use of SEO tools

(c) In addition to the RDCs, repositories, or websites used to publish the research data for a community, there are other services, aggregators, and so on that also index these datasets. We were interested to see the span of such services for the datasets of the KonsortSWD communities. Fig. 10, which shows the details on the research data availability across services, shows that DataCite[4], FDZ Datensuche[5] and BASE[6] are the top 3 services to offer datasets from participants' institutions.

In **(d)** we look at the data findability measures participants' institutions have adopted. The survey data show us that there is a variety of measures adopted, with the top 4 including providing open metadata for their datasets (available for search and harvest by other parties) for 77% of the participants, rich documentation of the data according to common standards (77%), use of key terms in important places on the website publishing the dataset (47%), or having websites optimized for mobile use (47%). Fig. 12 includes the rest of the measures.

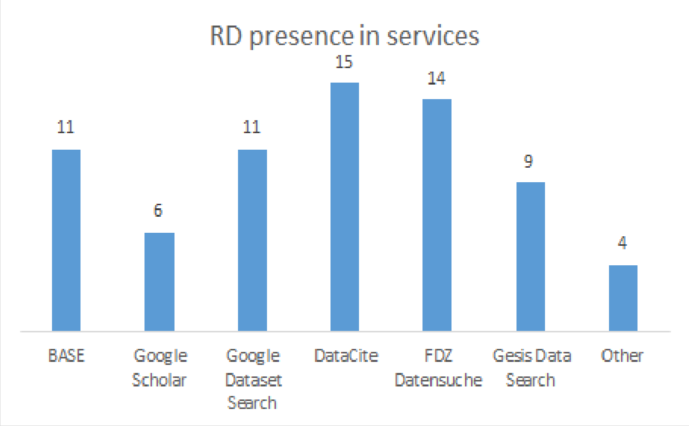


Figure 6 The presence of RD across services

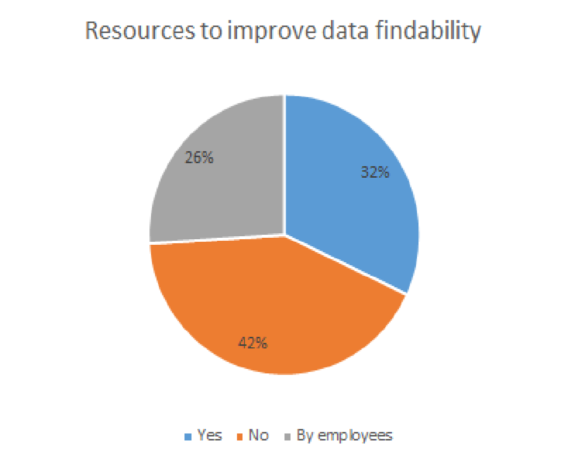


Figure 7 Resource allocation for data findability

As a final question, considering our measure's activities, we were interested in whether participants would be interested in improving their data findability and in what capacity. We noticed a different resource engagement for it; as shown in Fig. 11, roughly 1/3rd of the respondents reported that they have dedicated resources to data findability, whereas many more (42%) lack that. The remaining group (26%) reported that this is based on employees from the organization to pursue this alongside their current job description. On the other hand, the majority of the participants - nearly 70% - were interested in following our deliverables from the measure; 53% were interested in exchanging ideas and experiences with us, while 37% of them, in the top 3 positions, were interested in data findability recommendations. Fig. 13 shows the remaining options and choices for improving data findability.

[4] https://search.datacite.org/
[5] https://www.fdz-bildung.de/datenarchiv.php
[6] https://www.base-search.net/

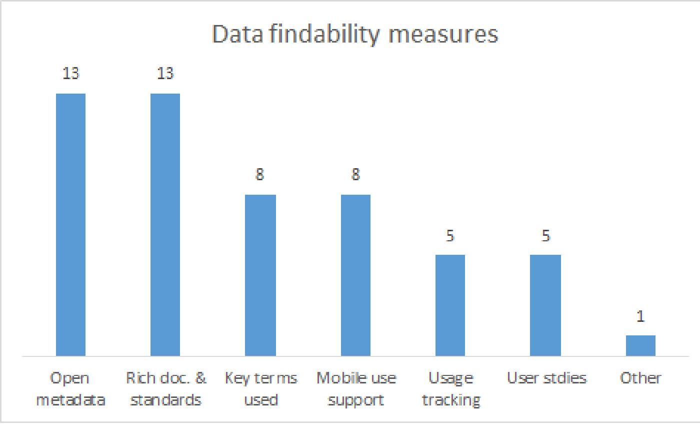


Figure 8 Application of data findability measures

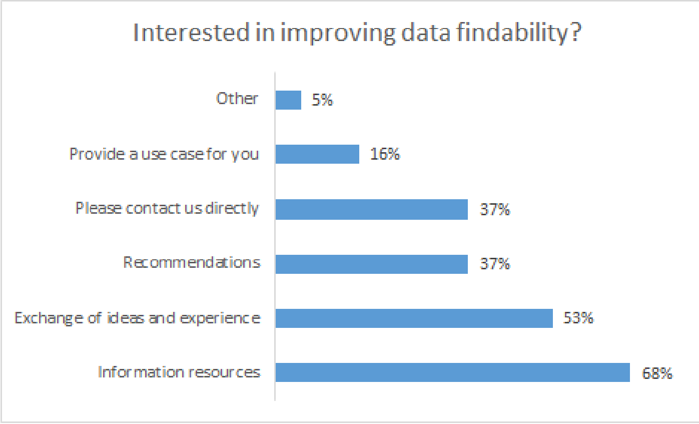


Figure 9 Participants' interest in improving data findability

We want to make a final remark on the survey results displayed in Fig. 8 and Fig. 11. On a first look, it could seem like there was a mistake in the labeling since both figures have identical groups in terms of quantity (42%, 32%, and 26%). We want to point out that these questions (and their corresponding charts) tackle two different questions in the survey and provide different options to answer them (Fig. 8: "Yes", "No", "Don't know"; Fig. 11: "Yes", "No", "By employee"). However, due to the response level, it might seem that they could be derived from the same set of data/responses in the survey.

1.3.3 Conclusion

Our goal for the first milestone of Measure 2 was to have a clearer picture of the state of RDCs in KonsortSWD with regards to data findability. One step in this direction was to adopt a quantitative technique and gather this information from as many participants as possible. The survey participation was satisfactory and enabled us to have a clear idea of the current activities these RDCs engage in the context of data findability.

88% of the respondents report an RDCor a data repository or a website for their data, and KonsortSWD is seen as a potential support to their data findability efforts (> 95%). Thus, we know which RDC has SEO reporting, and we have identified some candidates that want to get involved in prototyping the findings from Measure 2.

1.4 Interviews

The survey provided us with a more general picture of the state of data findability for KonsortSWD communities. In addition, we invited survey participants interested to discuss more details about their current data findability practices to do so via a set of more qualitative techniques - interviews. There were seven participants from 7 different RDCs that accepted our invitation for a more in-depth interview, the details of which we will provide next. We first start with the planning, then move on to the highlights of each interview session, before concluding with the insights from the interview.

1.4.1 Interview planning

The interviews consisted of 1-hour sessions, structured as follows:

- A 10-minute introduction to data findability, including information on how Measure 2 aims to improve it.

- A brief, 1 to 2 minutes of the introduction dedicated to the SISTRIX tool, including a report on some aspects that impact findability or the RDC or website that an institution uses to publish its (meta)data collections.

- The main part of the interview (reserved for 20 minutes) to discuss the specific findability aspects based on the survey feedback for an institution.

Every interview session was tailored to the individual participant. Based on the survey feedback, we prepared a set of questions to further understand their rationale for adopting certain practices, and, where applicable, provided recommendations that we thought would make a difference in their data findability approaches.

1.4.2 Interview sessions

In preparing for the interviews, the survey provided valuable input as it helped us understand the current data findability practices and prepare questions that we wanted to address during the interviews. Although the interviews were structured - we shared our questions before the interview, it wasn't unusual for the session to cover other related topics.

Due to the fact that the interviews contain information that can be considered sensitive for the corresponding institutions, we keep the details of the individual interviews out of the report. However, if you have a vested interest in reading these in full, please contact the report authors.

Instead, we will give a high-level overview on the topics discussed in the interviews. Please note that we did not discuss all these issues in all the interviews and that in some cases the interviews expanded beyond these topics.

Metadata practices One of the first points of discussion included the metadata standards. After all, when it comes to data findability, it is one of the prominent topics to discuss. Specifically, we were interested to know the metadata standards or elements in use (and the corresponding rationale), or the reasons for not adopting one yet for institutions where this was the case.

The most popular metadata elements include dataset *title*, *description*, and *meta* (machine-processable, used to provide some basic page descriptions). On the other hand, there was a variety of metadata standards either already adopted, such as the da|ra Metadata Schema, or being considered for adoption, such as Schema.org, the Dublin Core, or a few DDI vocabularies (with the most important attributes mapped to da|ra metadata schema).

Metadata standards, such as Schema.org or similar ones, could help some of the RDCs provide more information to the end users, including cases where users need to distinguish a page on a search engine's Search Result Page (SRP) vs the final (destination) page that contains the dataset details. This (accessing dataset from the SRP vs the dataset web page) is often not clear to the users, and was raised from some of the interviewees as an issu e.

We identified a few other practices that either relate to or are affected by the metadata standards and individual elements. In a few cases, the institutions include or make their dataset collections available through metadata search aggregators, such as da|ra and OpenAIRE. Moreover, specific to one of the communities represented by one interviewee, users link to a dataset in their publications, which eases the access and findability to the (cited) dataset. On a similar note, another interviewee mentioned the request to their researchers to link to datasets' webpages, but provided no details about it. Finally, we encountered the use of user studies as an approach that institutions employ to understand certain user needs. We expressed our interest to know more about these user studies (which we noted in the survey results) and the institution that conducts them indicated that they will share their findings with us.

SEO is another category that affects data searchability and findability. One part of the interview questions aimed to know the acknowledgement for or any eventual SEO adoption plans or strategies, including any reliance on it to achieve the aforementioned goals. To this end, institutions use SEO tools, such as Google Search Console, Matomo, and eTracker.

Another point from this context was the practice of using sitemaps to provide more information to a web crawler or ultimately a search engine. As for other practices discussed during the interviews, we encountered cases of using sitemaps (including DOI values) to this effect, and cases where they were either not perceived or explicitly created, and were instead provided by the tool that created their web page (such as a Content Management System, for example).

In some cases, due to lack of resources or know-how, institutions did not adopt such tools. In addition, data privacy concerns represented another aspect of not adopting SEO tools or related practices. There were institutions, however, with allocated resources for SEO - be it local or by outsourcing it to external partners. In these cases, we discussed their experiences with SEO. .

What were they interested to know about? During the interview sessions, the participants had their own questions, especially for topics or practices that were relatively new to them. On the topic of metadata standards, the question about Schema.org and how it could affect their data findability was quite frequent. The participants, generally, were quite familiar with the effort of the major search engines resulting in a schema to support structuring (meta)data on the web. As it can be expected, some of the questions on this topic involved aspects such as resources needed, the learning curve required to adopt Schema.org, etc.

As mentioned in *Interview planning*, we used one to two minutes of the presentation to present a report that covered several aspects that potentially affect the data findability of an RDC or an organization's website used to publish the (meta)data collections. For this purpose, we used the SISTRIX SEO tool. The questions that stemmed from this part of the presentation constitute the other set of questions we got during the interviews.

The questions were related to the SEO tools in general, and SISTRIX, our SEO tool of choice, in particular. As a result, we were able to explain some of the key features of the tool, including some of the reports that were generated for the participants' RDCs or websites. Moreover, the questions differed based on whether the participants were already using an SEO tool or not. Those who did would typically try to compare the features between the different tools, or even ask specific questions about SISTRIX ("does SISTRIX track signals from Twitter?". For your information, it does: Facebook, Google+, LinkedIn, and Pinterest are currently covered), while those who did not use such a tool yet, were interested to understand how it could become part of their SEO strategy, key features they could benefit from, and so on.

What did we recommend? Based on the feedback - both the information they provided or the information we were able to get through our SEO analysis for their RDC and websites, we were able to recommend the following actions to increase data findability:

- Conduct user tracking and using keywords to describe their data collections in their websites
- Become part of metasearch engines and registries, such as da|ra or Re3Data registry, for example.
- Metadata mark-up: Once Schema.org is adopted (mapping the dataset page's attributes to the schema), make sure you check its application via Google Console or a similar tool. One should use the available tools to make sure that the metadata markup is as correct as possible.
- Metadata to include: In addition to the title (which is the obvious one to include when publishing datasets), *keywords*, *description*, *name*, and *spatial information* should be some of the basic elements that could improve the dataset findability.
- On the "Title" metadata element: Include synonyms or acronyms in either the title or the description elements.
- Adopt systematic naming (title, and other metadata elements) is beneficial;
- Optimize for mobile users: An ever-growing user group that should be treated (optimized for and monitored). We need to note that some of the participants already reported optimization efforts for these users.
- Aim for the basic SEO aspects, especially since it does not incur such high costs and is appropriate for cases where no resources for data findability are available.
- Rely on the more popular domain as an initial step to increase the chances for more visibility of a given (meta)data collection, without even considering any other aspects of data findability (the proverbial "low hanging fruit" approach).

Collaborate with us to improve data searchability and findability practices Few of the interviewees expressed - to a different extent - the readiness to collaborate with us on the topic of data searchability. We have to consider, in any case, that they are generally low on resources they could dedicate towards this goal, but remain interested to consider our recommendations on this topic. All kinds of collaborations are vital to our mission and we would like to use these final sentences to give our heartfelt thanks to all the support and time we got from all the interviewees and many other members of KonsortSWD in pursuit of this measure.

1.4.3 Conclusion

The interview follow up helped us deepen our understanding of the initial data search and findability practices we identified previously via the survey. Based on some of the practices they adopt and the challenges they face, our impression is that interviewees are generally familiar with this topic. In this way, there is a common understanding of the role that metadata description (which elements to use), standards (Schema.org was of great interest), and practices (the use of sitemaps, for example) play in order for their users to find and understand their (meta)data collections.

SEO optimization and tools was another point of feedback; while they were, by and large, already relying on one (e.g., Google Search Console and Matomo), they are planning to adopt one, or are familiar with such solutions, data protection concerns and lack of resources dedicated to such operations was preventing them to fully utilize the benefits that SEO brings. In any case, even for those that already were using an SEO tool, most of the interviewees were quite interested in the results of their websites/RDCs generated via SISTRIX.

Some of the recommendations we provided include adopting metadata standards and providing rich metadata descriptions, applying SEO to their RDC's websites, relying on metadata aggregators, locating RDC's website under a more popular, institutional domain to increase user traffic to the RDC itself, to name but some of the broader categories. A final and valuable outcome of the interview was that we were able to identify partners to collaborate with on data search and findability practices as we move forward in this KonsortSWD measure.

1.5 Keywords analysis

Data discovery is often described as difficult or challenging, especially in comparison to literature search (Kern & Mathiak, 2015; Krämer et al. 2021). From observations (Krämer et al., 2021) and surveys (Gregory et al., 2020; Friedrich, 2020) we know that data discovery is a complex process, involving many steps. For e.g., 75% of researchers often use literature review for finding data, 41% use domain data repositories, 31% often use their personal networks; the top-3 sources researchers rely on during data search are literature, search engines, and disciplinary data repositories (Gregory et al., 2020). Web search is another of these steps; at 59%, it is highly relevant on its own, but also has a vital function as allowing researchers to move between the different parts of their data discovery process, and as a step was used by all participants for exactly that purpose (Krämer et al., 2021). This model of item search, as shown in Fig. 1, is also consistent with the log analysis of data search portals that shows a high number of known item searches (Kern & Hienert, 2018; Kern & Mathiak, 2015).

What researchers typically use when searching for data is most likely not a DOI. Mayo, Vision, & Hull (2016) report that 6% of data citations use DOI, despite this being the recommended practice since 2014 (Fenner, et al., 2019). Instead an informal name, a web link, an abbreviation, or the primary publication are used as proxies (Belter, 2014). According to Kratz & Strasser (2015), this is what most researchers perceive as "correct citation". In fact, many datasets, not curated within large repositories, do recommend citing the primary publication in an attempt to improve their bibliometric scoring, which may or may not include a workable link to the actual dataset, which, in turn, requires the users to fall back on informal names and other distinguishing features.

As research data is gaining in importance, so are the practices to help out with its different requirements. We just noted the recommendation for data identification via DOIs, but there are other, broader initiatives being adopted. FAIR principles (GO FAIR, n.d.) signify such an effort, and include a set of practices that help make research data ultimately easier to reuse. **Findability**, for example, entails a persistent identifier (a build up to the use of DOIs for data identification from before) assigned to the data, including metadata to describe it, and services that register or index it to make it searchable. In relation to Fig. 1, it becomes clear how different initiatives and practices are contributing to supporting different data requirements, in this case, data discovery or findability.

Research data is currently a very dynamic field of interest. However, the very recent reports on comparable issues posit that, although journals are now more likely to enforce data citation, they usually do not enforce following recommended data citation formats (Khan, Pink, & Thelwall, 2020). Also, researchers will continue to read papers from before 2020 for the foreseeable future. Likewise, social contacts will most likely not provide a DOI, but an informal name and maybe the name of an institution or a person attached to it.

An alternative way to attract users is to become part of the general knowledge of the community. Services like Zenodo[7] and OSF[8] have become part of the general knowledge of many scholars, as have some domain-specific

[7] https://zenodo.org/

[8] https://osf.io/

portals, such as PubMed[9], for their respective community. The original source of fame for these portals is not necessarily the quality of the data they provide; they are a package deal. Zenodo offers free web publication of research data, whereas PubMed has started out as a database of journal paper abstracts.

The most basic service that needs to be provided by the data discovery infrastructure is to ensure the findability of data via informal names and other standard and non-standard data citation types. Once a user has found data within a particular part of the infrastructure, they can then use this part to find more data through direct search on that platform, if the platform was sufficiently useful and memorable. In this paper, we propose a method to analyze the search practice of using such informal/non-standard names during data search.

1.5.1.1 Methodology

Using keywords that attract free traffic without clicking on an Ad, called organic keywords, in specific positions of web pages like the title, header, description, etc., is one of the most recommended practices to improve findability. Search engines like Google, use organic keywords as one of the key factors in ranking the webpages. Since a dataset is ultimately presented as a web page, we can use the same approach (organic keywords) to also improve its findability. In any case, what remains to be determined are the important keywords that we should use for this purpose, considering the specifics of the domain of choice - in our case, that of social sciences. Our research tries to answer precisely this aspect - identify the most important keywords users use to search data in the social sciences domain.

1.5.1.2 Data

We collected data from different sources, which include queries from two different institutions (GESIS and DZHW), provided by Google Search Console. In addition, we got a file that contains datasets mentions in papers from the Social Science Open Access Repository (SSOAR). We will refer to it as a citation collection from now on. This citation collection has the added value of providing context about the dataset reference. By context we mean information about what comes before and after the dataset reference in the research paper, which provides more insight in our analysis. Let's briefly discuss these sources, before summarizing them in Table 6.

1. GESIS search query logs: This collection was obtained from Leibniz Institute for Social Sciences (GESIS) contains queries taken in the time period from 01.01.2020 to 01.04.2021, captured via the Google Search Console .
2. DBK query logs: This collection includes queries from 01.01.2020 to 15.04.2021. It was obtained from the Data Catalog (DBK), which has been discontinued due to a security issue but still can be accessed from GESIS search.
3. DZHW query log: This collection was obtained from the German Centre for Higher Education Research and Science Studies (DZWH), one the KonsortSWD partners, whose part this research is being undertaken.

[9] https://pubmed.ncbi.nlm.nih.gov/

4. SSOAR Citation Collection: This collection contains dataset mentions and text excerpts that come before and after them in research papers. We obtained this collection from the Social Science Open Access Repository (SSOAR), exported from their Elasticsearch instance.

Data Source	Data type	Format	Collection size
GESIS https://www.GESIS.org/	Search queries	CSV	1.000 queries
DBK https://dbk.GESIS.org/	Search queries	CSV	1.000 queries
DZWH https://www.dzhw.eu/	Search queries	CSV	1.000 queries
SSOAR https://www.GESIS.org/ssoar/home	Dataset mentions from papers	JOSN	63426 items

Table 6. Data collection sources: Search query logs and data mentions

Each query log item contains the following information: query, clicks, impressions, CTR, and position information, whereas a dataset citation item contains many pieces of information about a (set of) paper(s), but we are interested on what comes before (before:text) and after (after:text) the dataset citation (citation:text) in the research paper, i.e., before:text, citation:text, and after:text. Both the citation collection and query logs are taken from social science-related sources and represent how users target the datasets to express their needs. This, in turn, makes them a great fit as experimental collections for improving the findability of datasets in this field.

1.5.1.3 Approach

In order to identify the most important keywords, we adopted two approaches: in the first one, we rely on keyword probability as means of clustering criteria, whereas we base our second approach on the Click Through Rate (CTR) from the selected data sources. Both the probability and CTR are measures that indicate how important a keyword is to the process of data discovery.

Keyword probability selection The probability of a keyword is its frequency in the query log divided by the number of queries in that log. Here, the citations from the citation collection are considered as queries, and this is based on the assumption that researchers usually start searching for a dataset after reading about it in some paper. This behavior has been observed quite often in (Krämer et al., 2021). Moreover, they also tend to use the same keywords mentioned in the paper to search for the dataset on the web.

CTR selection This technique makes use of a normalized version of CTR, which we call Comparable-CTR (CCTR), as a selection measure of important keywords. Keywords with higher CCTR are more important. It is worth noting that this technique can be applied on query logs but not on citations because CTR is related to the clicks and impressions, which are not available in the citation collection.

1.5.1.4 Approach implementation

After presenting our approach ideas, we next detail their implementation. We relied on Tableau[10] as our solution of choice for clustering, whereas we used Python for data wrangling. We start with the probability-based approach, and then move on to the one based on CCTR.

1.5.1.5 Clustering Using Keyword Probability

The goal of this method is to find the most frequent keywords in each query log. To do so, the log files are processed as follows:

1. For each query in the log file, the query is split into words.
2. Stop words are removed (both in English and German). Other languages are not considered.
3. Numbers and punctuations are removed.
4. For each keyword, the probability, number of clicks, and number of impressions is computed. The number of clicks of a keyword is the summation of all clicks of the queries in which the keyword appeared. The same thing applies to impressions. The probability of a keyword is the number of queries in which it appears divided by the number of queries (in this case 1000).
5. Cluster the keywords using K-means (Ralambondrainy, 1995) with Calinski-Harabasz (Halkidi, Batistakis, & Vazirgiannis, 2001) depending on their probabilities.

After applying these steps on the query logs, we got the results shown in Table 3. Fig.14 depicts the result of applying the analysis on GESIS query logs. There are three clusters from this analysis: cluster two, shown in green, contains the best keywords (listed in Table 3). The best keyword from this log is *scale*, whose probability is 0.042, i.e, this keyword appeared in 42 out of 1000 queries. Similarly, analyzing the DBK queries resulted in four clusters, from which the keywords in cluster 3 are those with the highest probabilities. Additionally, cluster 4 contains an outlier in the positive sense: the only keyword contained in cluster 4 is *fragebogen*, which has the highest probability (it has appeared 110 times in 1000 queries). We note this because we use this keyword for the SEO analysis later on (Fig. 5). Finally, for the DZHW query collection, the analysis resulted with 7 clusters. Among them 3 are outliers. These outliers are clusters 1, 2 and 3, and in addition to cluster 4, contain the best keywords which are of interest to us. In cluster 4, the keywords *fragebogen* and *questionnaire* have the exact probability and meaning, but *Frage* and *Question* are even more popular. It should also be mentioned that the number of clicks in the DZHW queries is very low, with the best keyword reaching 10 clicks; this does not have an effect on this method but will affect the results of the next method, which uses CTR as a feature.

[10] https://www.tableau.com/

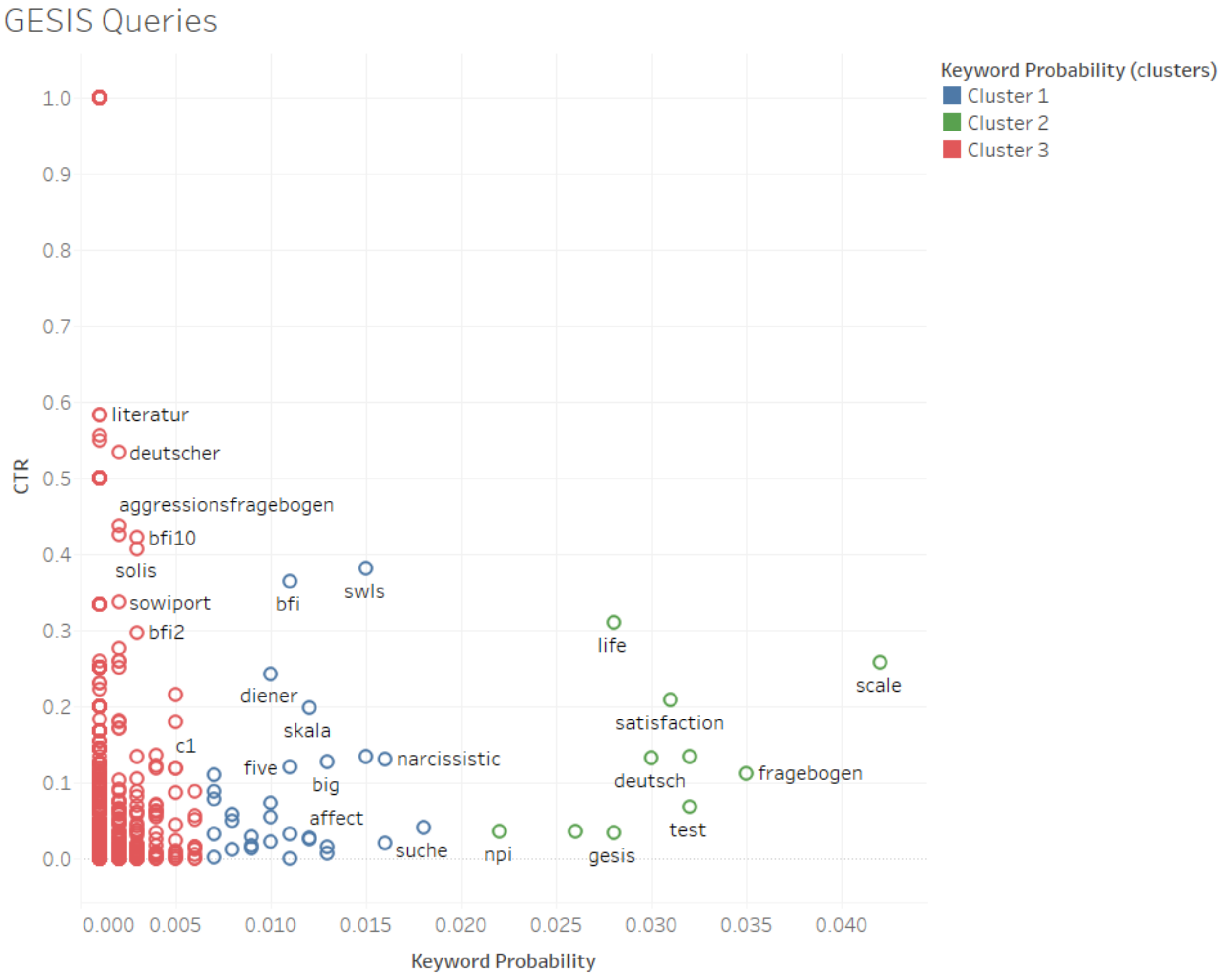


Figure 10 GESIS Query Analysis Using Probabilities

Finally, we want to make use of the dataset mentions in the research papers. As mentioned, it is normal for a researcher to use the context in which the dataset - or at least part of it - was mentioned in formulating a query during the search for the data. A query is formed from the before, term and after values from the citation collection. These queries were processed using the steps from 1-4 from the approach above, and the resulting keywords were ranked by their probabilities. After that, we took the keywords with a frequency of at least 500. Table 7 presents an excerpt of the long list of keywords that we got.

We want to make a brief clarification of some of the terms in this table. When it comes to the terms used for the analysis, there are three types of cases which we face in the data sources: some of the query words are acronyms of datasets in the repository; for e.g., "npi" stands for Narcissistic Personality Inventory; others are names of datasets or part of names; whereas the last type of query words are query specifications, such as questionnaire, which users use to qualify that they are looking for data or specify the kind of data. The last type of terms is very valuable as a general keyword for all pages that contain datasets.

Approach	Data collection	# of clusters	Selected cluster(s)	Top keywords	Probability/CCTR values
Keyword probability	GESIS search query logs	3	2	*scale*, *fragebogen*, *test*, *inventory*, *satisfaction*, *deutsch*, *GESIS*, *life*, *panas*, *npi*	0.022-0.042
	DBK	4	3, 4	*Fragebogen*, *GESIS*, *Deutschland*, *questionnaire*, *Eurobarometer*, *ddr*, *survey*, *English content*, *analysis*, *allbus*, *weimarer*, *republik*, *cooking*, *sa*, *codebuch*, *pdf*	0.014 - 0.112
	DZHW query log	7	1, 2, 3, 4	*question*, *frage*, *English*, *survey*, *fragebogen*, *questionnaire*, *interview*, *en*, *info*	0.027 - 0.123
	SSOAR Citation Collection	/	/	*Allbus*, *issp*, *survey*, data, evs, social, gles, study, european, analysis, ddr, panel, values, deutschland, ergebnisse, political, GESIS, used, questionnaire, fragebogen, question, frage, eurobarometer, studie, time, source, world, following, life, example, wahlen, cses, main, qualitative, religion, Beispiel, deutsche, content, educational, inequality, coding, online, interview, statistik	0.008 - 0.57
CCTR	GESIS search query logs	3	1	*swls*, *scale*, *satisfaction*, *personality*, *panas*, *narcissistic*, *life*, *inventory*, *deutsch*, *bfis*, *bfi10*, and *bfi*.	0.5 - 0.69
	DBK	3	1	*wehrstammbuch datenbank*, *GESIS*, *fragebogen*, *digitalisierung*, *dbk*	0.25 - 0.5
	DZHW query log	20	All clusters except cluster 20	*soziale*, *selektivität*, uni, mainz, *eoc*, *fahrzeugteile*, *graduated*, *fh*, *pr*, *kiel*, *covid19*, *maschinenwesen*, *de71*, *eua*, *cde*, *individuelle*,	0.8 - 0.09

				relationship, soziodemographischer, bildungsniveau, iab, remote desktop.	
	All Three datasets	2	1	*bfi, narcissistic, personality, inventory, deutsch, pana, swls, satisfaction, life, scale, bfi10, bfis, fragebogen, solis, datenbank, GESIS, suche, cobb, search, arbeitszufriedenheit, big, five, npi, test.*	0.109 - 0.125

Table 7 The data analysis for the two approaches using the dataset collections from Table 6

One way to combine the obtained results is to intersect the lists obtained from the query log collection with the ones obtained from the citation collection. Such intersection removes a lot of frequent keywords obtained from the citation collection because the number of keywords obtained from the logs is very small compared to the ones obtained from the citation collection. Another way which is more inclusive could be to take the union of all obtained sets.

1.5.1.6 Clustering Using CCTR

The first method uses keyword probabilities as the measure of importance, and this can be useful in many cases. But since we are dealing with datasets whose visits will not be so frequent, at least in the general case, keywords with high probability might not serve our purpose since frequency alone is not enough. We are looking for keywords with the highest number of clicks. That is why we make use of the CTR, which is another measure that indicates the importance of a keyword. It is defined as the number of clicks divided by the number of impressions (CTR = No. Clicks/No. Impressions), which takes values between 0 and 1. The higher the CTR of a keyword, the more important the keyword is. But according to this definition, a keyword that results in one impression and one click has the highest value (CTR = 1), whereas a keyword with 2000 impressions and 1000 clicks, for example, has a CTR of 0.5. Based on that, we cannot use CTR to compare keywords.

One idea to make CTR comparable is to normalize its values so as to make them comparable. Let the clicks be normalized into a range [0, a] and the impressions into a range [a, b], where b > a > 0; then the Comparable CTR (CCTR) is calculated by dividing the normalized clicks by the normalized impressions. The result will be in the [0, a] range. If we choose a = 1 and b = 2 then the CCTR will be in the range [0, 1] and it will have a value of 1 for keywords with the highest number of clicks and lowest number of impressions, a value of 0.5 for keywords with the highest number of clicks and highest number of impressions and a value of 0 for keywords with the lowest number of clicks.

The choice of a>0 protects us from the division by zero problem; we can then use CCTR to cluster the keywords and they separate nicely as will be shown shortly. This clustering approach is good if we want to know the keywords that have high clicks and low impressions which are of interest to us. The normalization that will be used is the min-max normalization which uses the following equation:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A \; - \; new\_min_A) \; + new\_min_A \text{ (1)}$$

Where v' is the new normalized value; $min_{A,}$ max_A are the lower and upper bounds of the old interval; and new_min_A , new_max_A are the lower and upper bounds of the new interval respectively.

Here are the steps

1. For each query in the log file, split the query into words,
2. Remove stop words
3. Remove numbers and punctuations
4. For each keyword, compute normalize the number of clicks into the range [0,1] and the number of impressions into the range [1, 2].
5. Compute the CCTR= NC/NI; where NC stands for normalized clicks and NI= normalized impressions.
6. Plot the keywords with the impressions as the x-axis and the clicks as the y-axis.
7. Cluster the keywords by their CCTR using K-means with Calinski-Harabasz.

Applying the above mentioned method on the three query logs yields the results listed in Table 3. As Fig. 15 shows, analyzing the GESIS queries using CCTR resulted in three clusters, of which the first one contains the best keywords. The DBK query log analysis also resulted in three clusters, of which cluster one contains the best keywords. Finally, the result of clustering the DZHW query logs using CCTR produced twenty clusters. Of the resulting clusters, cluster 20 contains the largest number of keywords whose CCTR is zero so its keywords are excluded. The remaining keywords are spread among the other clusters that contain one or two keywords in the best case.

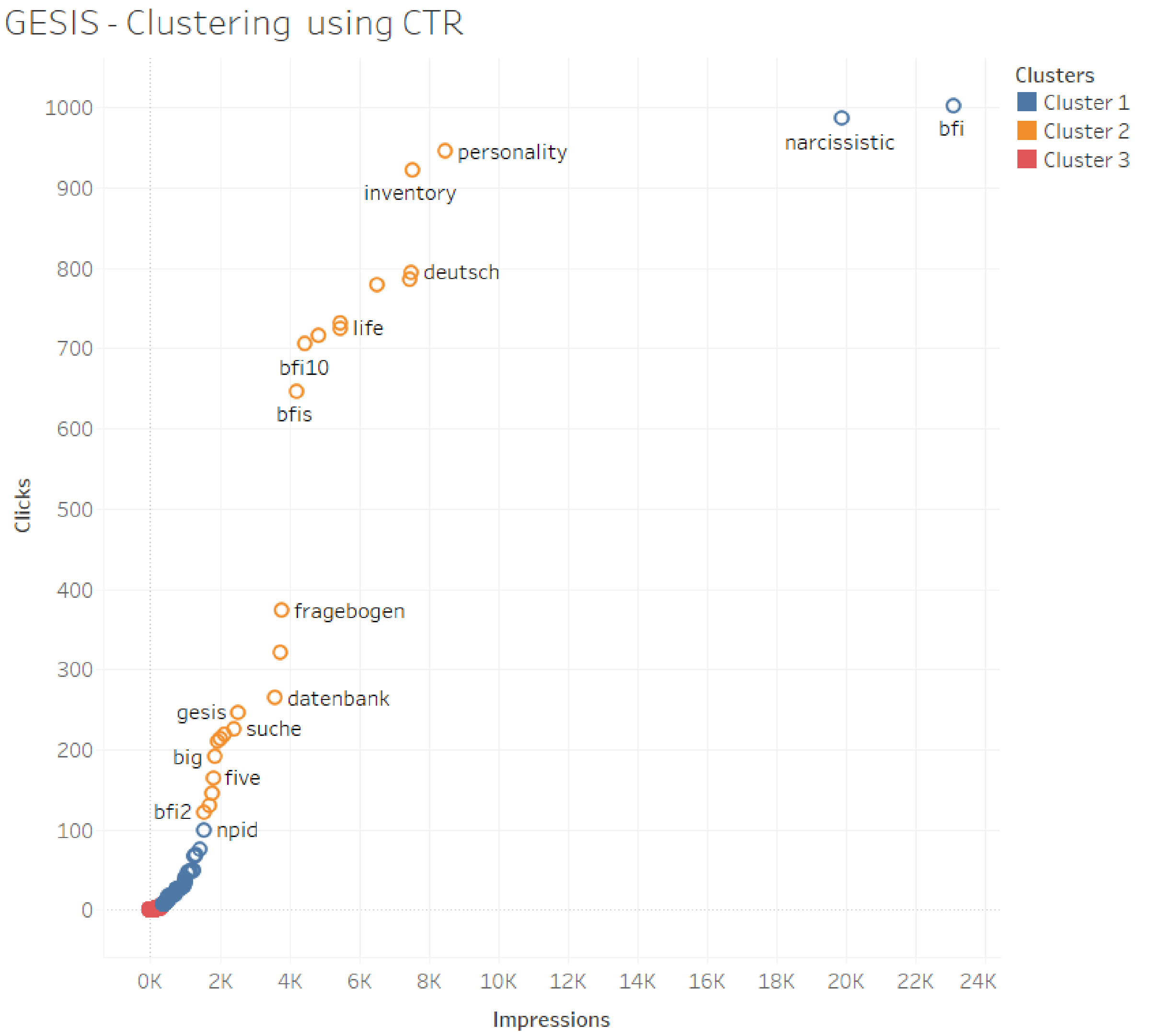


Figure 11 GESIS clusters identified with the CCTR analysis

The use of CCTR allows us to compare and cluster keywords not only from one query log collection, but also across different such collections. Fig. 12 shows the result of clustering all the keywords from the three query log collections combined together. The important keywords are contained in the first cluster and listed in Table 3.

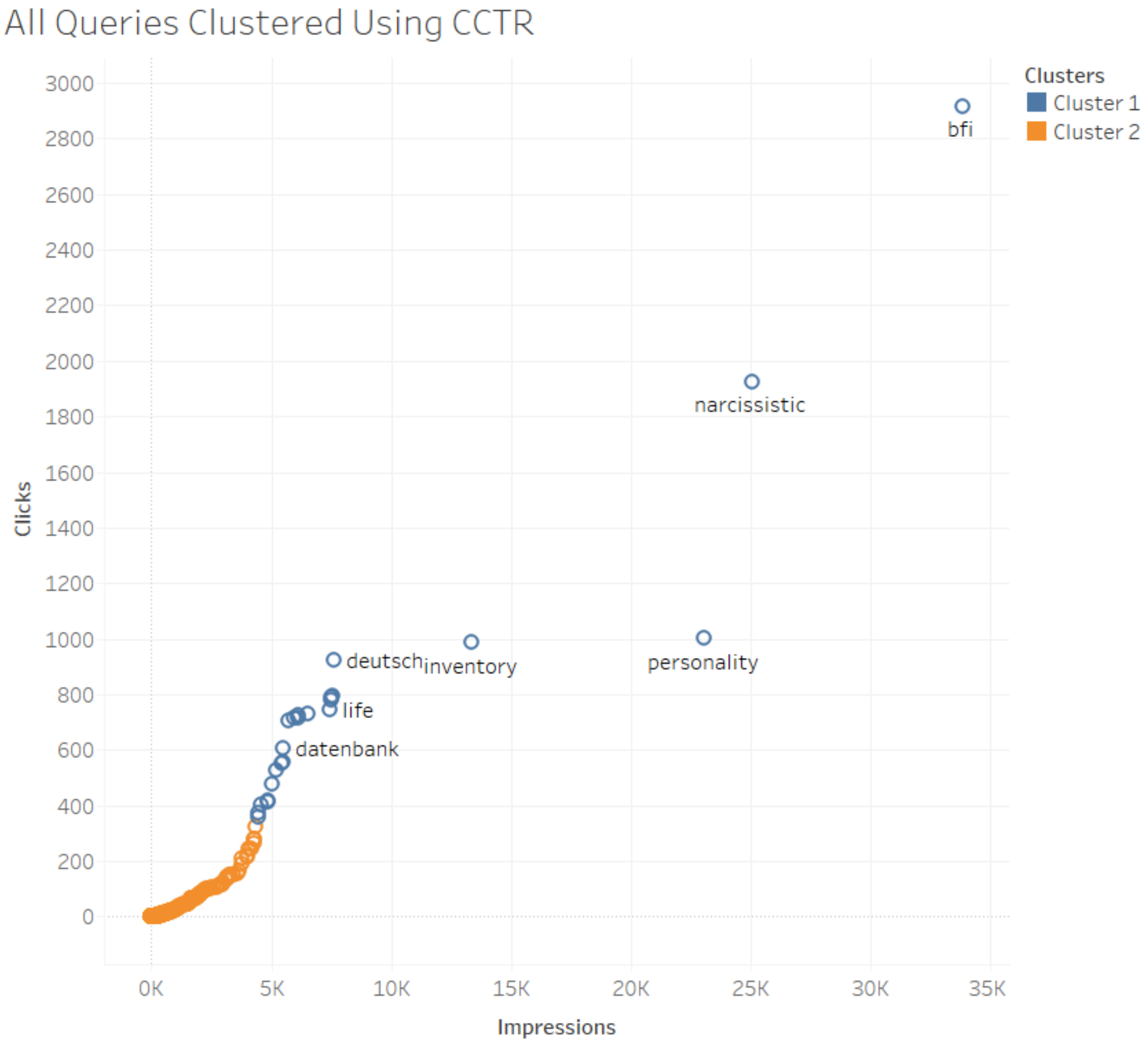
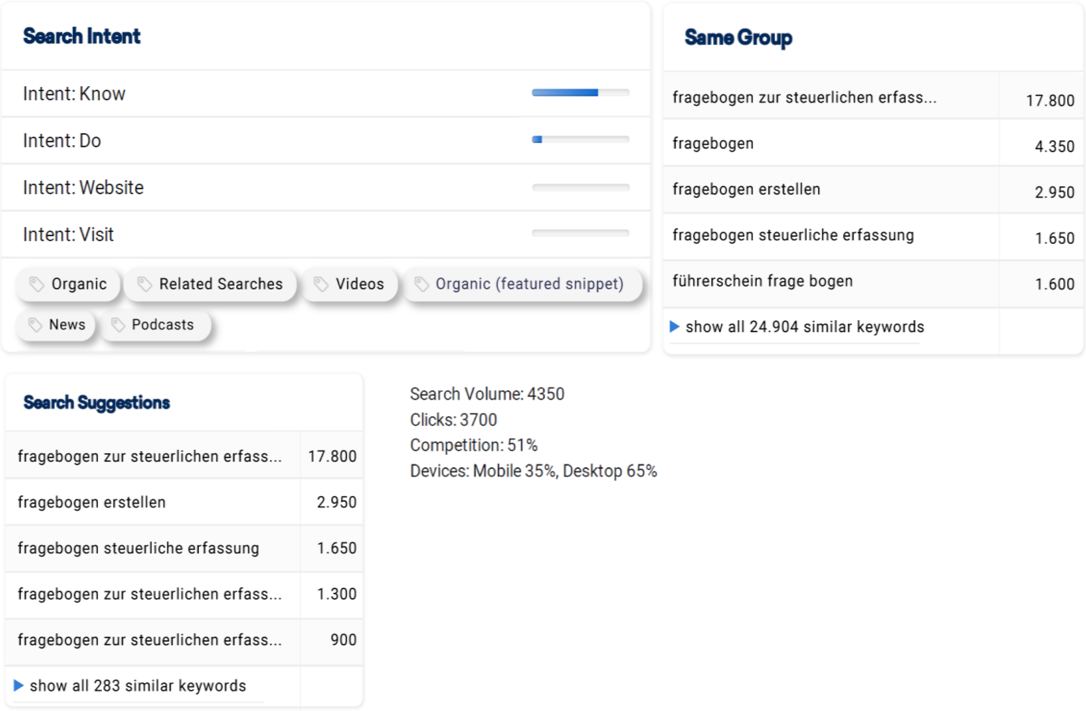


Figure 12 Keywords from different files clustered by the CCTR feature

There is one last thing to point out about the CCTR: with the choice of normalization intervals as above, the keywords which have the lowest number of clicks will have CCTR=0 regardless of the number of impressions. In practice this is acceptable because we are not interested in the keywords with the lowest number of clicks. But this could be a problem in the special case when all the keywords have the same number of clicks. In such a case, all keywords will have zero CCTR so it won't be practical to use CCTR with the above selected intervals for clustering. Instead, we can normalize into other intervals such as [1, 2] or [2, 3] for clicks and impressions, respectively. In this case, the CCTR will be in the range [0.3, 1]. The minimum value (0.3) will be associated with keywords that have the minimum number of clicks (1) and the maximum number of impressions (3), while the maximum value (1) will be associated with the keywords with the highest number of clicks and the lowest number of impressions. This works as well but for the sake of this work we stick to the choice of intervals that results in the CCTR being in the range [0,1] so that we can show the similarity to CTR.

1.5.2 Evaluation using SEO tools

Now that we have a set of keywords to start with, we can rely on Search Engine Optimization (SEO) tools for our evaluation. The keywords can be fed into an SEO tool, such as SISTRIX[11], to give us information about the most useful keywords and how we can use them. For example, *fragebogen* is a common keyword in the keyword lists above. If we analyze it using SISTRIX we can get the report in Fig. 13.



Search Intent

Intent: Know
Intent: Do
Intent: Website
Intent: Visit

Same Group

fragebogen zur steuerlichen erfass...	17.800
fragebogen	4.350
fragebogen erstellen	2.950
fragebogen steuerliche erfassung	1.650
führerschein frage bogen	1.600
▶ show all 24.904 similar keywords	

Search Suggestions

fragebogen zur steuerlichen erfass...	17.800
fragebogen erstellen	2.950
fragebogen steuerliche erfassung	1.650
fragebogen zur steuerlichen erfass...	1.300
fragebogen zur steuerlichen erfass...	900
▶ show all 283 similar keywords	

Search Volume: 4350
Clicks: 3700
Competition: 51%
Devices: Mobile 35%, Desktop 65%

Figure 13 SISTRIX report for "Fragebogen" as a keyword

This SISTRIX report contains four sections:

- The "Search Intent" section shows that users who use this keyword have the intention to know about something and only a small portion of them have the intent to get or download something;
- The "Same Group" section contains queries that contain the keyword. This feature is good if long tail keywords are of interest.
- The "Search Suggestions" section suggests queries that may or may not contain the keyword but are semantically related to it.
- The fourth section contains miscellaneous information about the keyword. The search volume indicates that there are 4350 monthly queries that use this keyword in the specified country (Germany, in our case); 3700 clicks organic clicks for the keyword. Moreover, the "Competition" field of the report (51% in this case) indicates how difficult it is to rank for the given keyword. The higher the competition value,

[11] https://www.sistrix.com/

the more difficult it is to rank for that keyword. Finally, the "Devices" field indicates that 65% of the users used a desktop to do the search while 35% used a mobile device. There is other information that SISTRIX provides but we did not list them for brevity.

1.5.3 Conclusion

The difficulties involved in collecting research data and the need for such data for result reproducibility have brought attention to the data findability problem. In this work, we tackled this problem from the keyword point of view. On one side, using relevant keywords in formulating search queries plays a vital role in reaching the targeted data. From the other side, using relevant keywords in specific positions of the webpage is one recommended search engine optimization technique that helps improve the page's rank on the search engine result page.

Since log files generated by search engines are reliable and easy to get, we wanted to analyze some of the features in those logs and identify the most popular terms to search the data collections with. To do that we presented two different techniques to select important keywords from search logs and citations - based on the popularity of keywords and the CTR. We applied these techniques on 4 datasets from two different institutions, all from the domain of social sciences, with three datasets containing query logs (1,000 per dataset), and one containing dataset citations (63426 in total). Applying our techniques, we were able to understand the user search practices, and not only do this for the individual collections (and their corresponding institutions), but also explore these practices across collections (and their corresponding institutions). We achieved this by using a normalized version of CTR, which enabled us to compare the CTR values for the available log collections. By relying on an SEO tool, we were able to confirm the findings from our analysis, i.e., the popularity of the same keywords we found out during the analysis process.

The keywords obtained in this paper can be used to improve the findability of research data in social sciences. But the two techniques introduced are general and can be applied in a similar fashion to logs or research papers from any domain.

1.6 Conclusions and impact on the strategy

After broadly reviewing the current data searchability and findability practices, we focused on the KonsortSWD communities, employing a survey, and following up with a set of interviews. To help us prepare for gathering the requirements from this community and complete all the aspects, we relied on SEO as an important and very practical approach to this process. In this section, we summarize the findings we derived from all the employed techniques and outline how they are important to the KonsortSWD goals and the Milestone 2 Strategy report.

1.6.1 SISTRIX **SEO section**

- **KonsortSWD partners are not competitors to each other. However, some partners have multiple subdomains, which are competitors to themselves.**

  This is an important insight that will influence our general strategy as outlined in the **Section 2.1** (Centralized or Decentralized). What it actually means is that it can improve the visibility of one RDC's data, without harming that of others.

- **Main domains are generally more visible than their subdomains, but the subdomains inherit visibility nonetheless.**

  This is another important take-away, which we will use in the Milestone 3 report.

- **All the RDC we looked at have a strong focus on German-speaking countries as their primary user group.**

  This result is expected and needs no adjustment at this point.

- **The SISTRIX visibility index corresponds roughly to what we already know from user tracking of the respective domains. It can therefore be used as a tool to track SEO improvements over time.**

  We recommend using Sistrix as part of the monitoring infrastructure to be set up in Milestone 5.

1.6.2 Survey

- **The survey participation was satisfactory and enabled us to have a clear idea of the current activities these RDCs engage in the context of data findability.**

  We take this as encouragement to re-run the survey on a regular basis as part of the monitoring strategy.

- **88% of the respondents report an RDC or a data repository or a website for their data, and KonsortSWD is seen as a potential support to their data findability efforts (> 95%).**

  Web findability is a relevant issue for the majority of RDC and KonsortSWD is a welcome cooperation partner on this issue. This is an important result which makes many of our dissemination efforts much easier.

- **We identified the SEO reporting each RDC has;**

  This is another important result that influences our decision to re-run the survey as part of the monitoring strategy.

- **We identified candidates that want to get involved in prototyping the findings from Measure 2.**

  This is very important for the reference implementation(Milestone 5) and makes us optimistic that we can find good partners.

1.6.3 Interview

- **Based on some of the practices they adopt and some of the challenges they face, interviewees are generally familiar with the topic of data searchability and findability on the Web.**
- **There is a common understanding of the role that metadata description (which elements to use), standards (Schema.org was of great interest), and practices (the use of sitemaps, for example) play in order for their users to find and understand their (meta)data collections.**

  More information for stakeholders enables them to make the relevant changes themselves. Next, we will need to monitor whether or not this is a suitable way to make effective changes.

- **SEO optimization and tools was another point of feedback; while they were, by and large, already relying on (e.g., Google Search Console and Matomo), planning to, or familiar with, data protection concerns and lack of resources dedicated to such operations was preventing them to fully utilize the benefits that SEO brings.**

This particular issue encouraged us to include this into the general information package.

1.6.4 Keywords

The results from the keyword study deeply influenced the recommendations to be found in the Milestone 3 report **section 3.4**. It also shows the added benefits of using a decentralized monitoring system and exchanging data with one another. We hope that with the planned prototype (cf. section XY, Reference Implementation) and more data donations, more insights like this can be derived from this measure in the future.

2 Milestone 2: Strategy to increase the findability of KonsortSWD data on the web

Our work for Milestone 1 gave us a good understanding of the status - practices, motivations, existing infrastructure capacities, etc. - of searchability and findability of KonsortSWD data on the web. The methodology we adopted for this included a few different techniques that enabled us to engage with this community to different levels (via survey and interview), monitor, and analyze their searchability and findability measures in practice (via SEO and keyword analysis).

Based on this, we next wanted to seek ways to improve the data searchability and findability for the KonsortSWD members. We wanted to achieve this by means of a strategy that individual members could adopt as a necessary part of their practices. This is the focus of Milestone 2.

Based on the interactions we had with the KonsortSWD members so far, and the general trends and practices we considered so far, we identified 3 main components for a searchability and findability strategy. The strategy answers 3 corresponding questions:

1) How to balance and manage local vs global SEO requirements?
2) How to improve the dissemination and marketing of the measures for a maximal effect in the community (members' groups)?
3) What are some practical recommendations that participants could adopt as part of their searchability and findability practices? (This also links into Milestone 3).

2.1 Centralized or Decentralized approaches to SEO

2.1.1 Introduction

SEO plays an important role in supporting searchability and findability efforts for an organization. As already mentioned, we rely on SEO for some of the analysis in the report. Having in mind the distributed aspect of the members - not only part of different institutions, but also having their distinctive requirements within them, it becomes important to consider the way to manage or balance this situation. In particular, this means, whether to rely on centralized services, such as da|ra or DataCite and improve their SEO ranking, or whether to improve the individual participants.

Before we can discuss how to achieve this balance and the recommendations in our strategy, we need to define what exactly we mean by centralized and decentralized in the SEO context. In doing so, let's start with discussing the considerations of managing business operations in a global vs decentralized approach. In a context of multinational enterprises, Prahalad & Doz (1987, as cited in Chung, 2018) treated the balance between having a central control over the dispersed units, and, at the same time, responding to the (often unique) requirements that those units might have. They pointed out that successful companies have a clear approach - either centralized or decentralized.

When it comes to SEO approaches, they also can be seen from the centralized vs decentralized perspective, and Macdonald (2011) discussed the implications of such a model. He predicts a trend for large companies moving their SEO operations (from a centralized) towards a more decentralized model. In valuing the importance of local markets, it will contribute to a model shift that is positioned to better respond to the requirements from these markets. The author also mentions the benefits that the centralization model brings, but these relate to other operations, such as metrics development, reporting, technology acquisition, etc. that still are more applicable from a central standpoint (for e.g., if we want to know how well we are doing at the company level, having centralized metrics would enable such analysis).

Having in mind the decentralization trends, we want to address its implications for SEO in the KonsortSWD context. Specifically, what model are we looking for for its communities? How does a practice, such as the one reported for multinational enterprises, transfer to the SEO strategies for these communities? In any case, we consider this model just as an initial framework, and explore how an SEO strategy could take shape.

Biribicchi (2020) argues against decentralization, but he considers this from the perspective of a single company. In this case, the options are between a single and multiple websites. In any case, categories that he uses to conduct the analysis are relevant to our case as well. We have to note, however, that our case in KonsortSWD is slightly different, as there is no classical "centralized vs decentralized" model as seen in the previous references since we are dealing with multiple organizations. Thus, there is not a single organization - let's call it KonsortSWD - that is distributed across several units. Instead, we can see all the RDCs in the project as individual, centralized operations. Finally, our survey and interviews provide the most important input to analyze the centralized and decentralized models. Let's see how such an input fares across a few categories of interest for the analysis.

The occurring aspects for the two models minimally include (to not say, boil down to) the following:

- **Effort** This aspect includes the effort/involvement required to apply certain SEO practices;
- **Cost** Often directly related to the previous category, but could also imply additional information(even if there is no information about effort available, for example), based on inherent - often technical - costs of the model (if they are present).
- **Expressivity** How expressive can we be in terms of SEO for the RDCs of interest? This category presents such concerns, that is the availability to conduct SEO tasks, such as reports, monitoring, optimizations, etc., or tasks required by any of the other categories (for example, the effort required could include tasks that are of a certain specificity to apply across all or only on certain RDCs).

2.1.2 Results

Based on the input from the Milestone 1 report and additional sources, we were able to identify several factors that impact the adoption of one or another SEO model. These factors emerged from the discussions we had in the interviews, within the Task Area and with external players. Table 8 lists them, along with a brief discussion about their suitability for each mode (centralized or decentralized). Please note that this is a snapshot and some or all of the contributing factors might change in the future.

Factor	Suitability discussion	Centralized vs. Decentralized
Status Quo	The current status quo is de facto a decentralized model. Centralized services, such as da|ra and DataCite, as well as services outside of KonsortSWD, such as Google Dataset Search and openAIRE, just to name a few, tend to have fewer users (as far as we can tell by observing web traffic) and generally rank lower than original content as provided by the RDC.	Decentralized
No access to most of the centralized systems	Most of the centralized systems are outside of our influence. The only exception is da|ra.	Decentralized
SEO optimization costs	While it is relatively easy to lower general web visibility, increasing it is quite difficult. Generally, it is easier to promote smaller amounts of high quality data, with a lot of inlinks, than many uniform pages with similar content. In case of da|ra, the only central search service that we have direct access to, these costs would be very high and require severe restructuring of its web structure. Moreover, da|ra was never intended or designed to serve such a function.	Decentralized
Ongoing Activities	The RDCs in KonsortSWD apply different data findability measures. This includes providing the metadata openly, rich documentation and standards, key terms for the data description, etc. As also confirmed by our interview, the RDC are at different stages of data visibility, which, as a model, suits best the decentralized approach. Where a centralized model could fit better is for cases where the communities (thus, RDCs) lack even the first steps in terms of data visibility practices. Having a set of recommendations, such as the use of sitemaps, for example, although developed from a central perspective, could apply (to a different extent) to some of the RDCs.	Decentralized
RDC identity	Many of the RDC have expressed to us a desire to have individual visibility for their data, which is in accordance with their mission to promote the datasets for which they are responsible. Organizations, which do not wish to have an individual platform for their data, have many options to store their data with suitable RDC or other platforms. As such, we believe this sentiment can be generalized to the RDC, we did not interview individually.	Decentralized
The general structure of RatSWD and KonsortSWD	RatSWD itself and KonsortSWD follow a decentralized strategy.	Decentralized

Monitoring and reporting	Similarly to the "SEO optimization" category, a decentralized monitoring and reporting on an SEO campaign or strategy requires more effort - and potentially costs. This manifests itself in the form of individual monitoring (SEO) tools, or the generation of tailored reports for different RDCs. This approach, as seen before, is better suited to the individual target groups, RDCs in this case. A centralized approach, on the other hand, would reduce some of the effort, thus costs, mentioned before, but it might not allow the same "expressivity" to generate more detailed monitoring and reporting capabilities for the different RDCs.	Centralized

Table 8 Weighing the pros and cons: Centralized vs. Decentralized model for SEO

This rather overwhelming result in favor of a decentralized approach may be counter-intuitive to readers familiar with the current situation with literature. For literature, the current status quo is a centralized SEO approach. Centralized search engines, such as Google Scholar or Semantic Scholar[12], or those dedicated to certain disciplines, such as the Computer Science Bibliography[13], are extremely important for the visibility of individual papers. However, as we have shown, the situation is quite different for the research data landscape of KonsortSWD, where the decentralized model and the more cost-effective SEO practice is currently the status quo.

2.2 Suggested Actions to increase the findability of KonsortSWD research data on the web

In this section, we outline some of the planned activities for the remaining two years within the measure. We will follow the suggested milestones in the proposal but provide more details on how to implement them as well as describe supplementary activities and why we believe they fit in well with the strategy of KonsortSWD.

The milestones are as follows

- We plan to develop a reference implementation to explore and test the current and emerging SEO data practices; (Milestone 4)
- We establish KonsortSWD-wide **monitoring** to allow us to identify the adopt of current SEO data practices and the identification of new ones, (milestones 5 and 6)
- **Dissemination and marketing** activities: Reach all the targeted RDCs with the SEO data recommendations, as well as promote these findings in a broader context; (as part of milestone 7)

2.2.1 Reference Implementation

There are different ways for us to provide feedback to the KonsortSWD communities. Based on the first three milestones, we are able to share the information with them via a (published) report, such as the one you are reading, or engage a bit closely with them and provide them with a reference case that includes the findings we want to disseminate. While the former presents a nice way to formally describe the different aspects of these findings, its scope and delivery form (text) often limits the maximum impact with the recipients. The latter, on the other hand,

[12] https://www.semanticscholar.org/
[13] https://dblp.org/

is more efficient in providing a concrete implementation reference, and quite suitable to SEO practicants, which are the ones implementing any data SEO recommendations we disseminate.

When it comes to realizing the reference implementation described above, we plan to apply data SEO enhancements directly at the community repositories, which suits our objective better. The distributed nature of the RDCs - and the SEO model adopted in this project, as discussed in the previous section of the "Strategy", and the variety of requirements that comes from the different RDCs, needs a close and individual collaboration in order to successfully explore and prototype the data SEO practices.

Our choice for the reference implementation implies that we identify several partners and help them with their data SEO. In addition, we will be in a position to include (case) specific recommendations, including implementation aspects, wherever possible. The latter is quite important especially for the institutions with no resources dedicated to SEO operations.

In our survey and later with the interview sessions, we made sure to include a question and discussion point for the participants about their potential availability to participate in activities such as the ones that pertain to the reference implementation. Thus, we already have an initial list we can start with for this task. However, we plan to contact the survey participants again with a proposal to collaborate on improving their SEO, especially once few such collaboration examples are available. What we have noticed in our work so far is that practical examples, including case implementations (even better if it is from similar/relevant RDCs) or SEO tool demonstrations, always peak the interest of KonsortSWD communities, so we hope that to be the case with this task, too.

2.2.2 KonsortSWD-wide Monitoring

As with any other case, SEO practices are not static by any means. Namely, as new practices emerge and are disseminated to the broader audience, communities tend to adopt them. While this happens at a different pace for different communities, or even members within the same community, it is important to be aware and plan accordingly. What this tells us is that SEO practices need continuous monitoring, that is, studies which we might have done in the past could be (even if slightly) outdated to be used for a longer decision making in the future.

Thus, one of the aspects of disseminating and marketing data SEO practices in this measure is that of continuous monitoring. Here, we plan to (1) re-run the survey - and tentatively the interviews - we did as part of the study from Milestone 1 is the task for this objective. This should happen in regular intervals.

Another, complementary step is (2) re-applying SISTRIX to the KonsortSWD RDCs, which proved efficient for surveying the SEO state of the RDCs - of interest both to us and to the participants themselves. In our experience, this set of techniques has proved as an effective approach to conduct the monitoring required for this part of the dissemination and marketing efforts of the SEO strategy.

Both of these monitoring activities will be documented in M6, the evaluation of the impact of the implemented visibility strategies.

2.2.3 Dissemination and marketing activities

In order to ensure that the findings and recommendations that derive from this measure are disseminated to the target audience, we plan on a few techniques. One we would like to start this part with is a flier that would include some of the more common - but practical - SEO practices community members can adopt. We already have enough insight from the data gathered in the course of the measure, especially the one collected via the interviews, to start working on a first draft of the flier. Although a digital version of the flier would reach broader audiences faster, we are also considering the physical fliers - in forms of popular stickers, for example - as another (dissemination) alternative.

Our dissemination and marketing activities are also manifested by our participation in relevant organizations and events. We are quite active in the Nationale Forschungsdaten Infrastruktur[14] (NFDI), and regularly participate in the events organized by it. Specifically, we are active in the NFDI section on Metadata, Findability and Provenance. One such event, InfraTalk, for example, is planned for January 2022, which we plan to attend.

Outside of the NFDI, we already actively participate in different data search and findability forums (the RDA working groups and interest groups are but one such event; see for example the Data Discovery Paradigms Interest Group[15]) as means to stay abreast of relevant initiatives and recommendations, including concrete deliverables organizations can adopt in order to improve their data searchability and findability. Participating in such events also implies being part of the organization, panel discussions, and presentations of new research. In the context of this activity, we plan to continue - and possibly extend our outreach with - our participation in these events.

Ongoing discoverability activities include the GO Fair Implementation Network Discovery and RDA Data Granularity. These network activities are to be continued and expanded. The goal is to actively participate in the recommendations on standards and best practices and to represent the interests of KonsortSWD.

2.3 Hands-on recommendations for the participants

From both the survey and the interviews, we know that there are several gaps for the individual RDC in terms of both knowledge of what to do and resources to do it. Due to lack of time, some methodologies cannot be implemented in some RDCs. Furthermore, the use of a tracker for user tracking or for analyzing the findability is highly beneficial and very easy to implement, but not all RDC utilize them. With some easy steps, data can be collected and the analysis of the information can bring new insights. In this step, some also lack knowledge about different trackers, e.g. some RDCs cite data privacy concerns when not utilizing web tracking, which is true for some systems, but not for all. With some clarification, the uncertainties could be eliminated.

On another note, in the cases where the collection is on another domain, moving it to the more visited domains could be a good progress to increase the chances for more visibility of a given (meta)data collection. More issues will be introduced below. This is something we identified during our SEO analysis for most of the interviewees' websites or RDCs.

Some hands-on recommendations for the participants, including optimizing coverage, impressions and clicks of research data (SEO), are now in order. In addition to the knowledge of SEO techniques, success requires above all a precise knowledge of the technical details of the portal to be optimized, its target group and unique selling points. SEO specialists must work closely with portal managers and technicians. Especially in maintenance mode, portals are often thinly staffed and have little capacity for SEO (as seen in the "Survey" section). Active projects, on the other hand, are moving targets. With every change in user interface potentially comes changes to visibility.

[14] https://www.nfdi.de/

[15] https://www.rd-alliance.org/groups/data-discovery-paradigms-ig

Depending on resources and strategic planning, the following measures are therefore planned for the portals. The goal in each case is the individual optimization of coverage, impressions and clicks at the portal (or website) level.

2.3.1 On-site Monitoring

Few of the first measures that we propose in this section include the creation of sitemaps, integration of SEO tools, as well as those for user tracking. Let's briefly discuss the rationale for each one next.

After a relatively long presence, many argue about the value of sitemaps. In any case, since we are trying to improve the findability of data collections indexed by search engines on the web, having an explicit list of pages for search engine crawlers to process remains an important first step towards higher data findability. Moreover, for the cases where the website representing the dataset is relatively new or has minimal number of backlinks, using sitemaps remains an important step to allow search engines index them (Cojocariu, n.d.).

As of January 2021, Google search engine has 87% of the market (Statista, n.d.). Thus, it becomes important for any data collection on the Web to be aware of its standing in relation to this search engine. Luckily, there is a service - free of charge, no less - that anyone can use to this end - the Google Search Console[16]. This service provides a set of features that help monitor and improve the standing of your website. Being that it is a feature-rich, free service, provided by the biggest search engine itself, this is another important recommendation that allows efficient monitoring capabilities to the RDCs in the KonsortSWD community.

User tracking is another measure we recommend, as it allows tracking different usage aspects of a (data collection) website. This includes aspects such as where the users are coming from, which pages are generating the most traffic, how - via which patterns - users navigate a website, and so on. The level of tracking can also include individual users for websites that have registered users. For all of these features - and more - there are tools that one can adopt to conduct user tracking. One such example tool is Google Analytics[17], which is but one of the options and a powerful, often free of charge, tool that can be readily used by interested parties.

In a final remark, we would like to make a note on data privacy concerns. Some of the measures we propose in this section, especially the integration of tools and user tracking, need a more careful consideration from the data privacy aspect, as they can collect user behavioral information. In any case, we need to stress the fact that all such information will be collected based on the General Data Protection Regulation (GDPR) mandates. Moreover, the focus and the expected benefits from such insights will be planned in the context of improving data searchability and findability in the KonsortSWD communities.

2.3.2 Using tools to find and fix errors

In the previous part, we already introduced two example tools to support SEO operations. Although popular, and developed by the company behind the most popular search engine in the market today (Google), there are many alternatives to consider. This task, ultimately, is for the KonsortSWD community members to make, and we just provide a few of our preferred tools as a starting point for them.

[16] https://search.google.com/search-console/

[17] https://analytics.google.com/analytics/

Some of the tools such as the Search Console and Analytics from Google offer features that help webmasters prepare the website for a better overall user experience. This includes different considerations for the type of device used for access (desktop or mobile), the loading time, any (HTML) errors on the site, and so on.

During our interview, we had a chance to apply our SEO tool of choice and analyze the websites of 7 RDCs. One aspect that we always tried to show to them was the evaluation that SISTRIX provided for the desktop vs mobile usabilities. Considering the considerable number of users accessing data collections in these websites from their mobile devices, implies a need to treat it as an important requirement. Improved mobile usability is particularly important, as crawlers are on the move as mobile devices and measure usability in the process. Google Search Console, for example, offers concrete suggestions that can be implemented with more or less effort. As an example, approximately 20% of the users come to GESIS pages with mobile devices. Manual tests and, if necessary, changes to the interface increase usability for users in addition to the technical measurements.

Finally, aspects such as load times of the pages can have an impact on the overall user experience for a website. This is also an aspect that (SEO) tools can identify and help website administrators address on time. On another note, in order to make sure that a website is crawled and indexed by search engines, we recommend relying on the robot.txt files. Making sure that it represents the part of the website that we want to eventually show up in the search engine results page, while also removing any potential sensitive or pages that do not provide much value for a website (login form, for example).

2.3.3 Inlinks and Outlinks

We reserved a final set of recommendations for the role of links both within (inlinks) and to (outlinks) external resources. For the former case, few recommendations such as linking to the more important pages of the website, or using text - potentially the keywords of the dataset - as part of or around the links, could be applied (What Are Internal Links (Inlinks & Outlinks)?, n.d.).

On the other hand, linking to resources outside of a website for a data collection is also important and pursued as an approach by many organizations, and not only when publishing dataset collections. As they are usually non-profit organizations, RDC and individual RDC offerings, for example, have opportunities to post links to themselves in Wikipedia. Wikipedia has a very high PageRank and also provides the information for the info box that is displayed on the right for many search queries. This means that several search paths are used at the same time. On the one hand, the ranking of the classic web search results is improved, and users can be reached directly via the info box. Other approaches to external linking include listing the data collection website on data registries, such as re3data.

2.3.4 Other issues

In the SISTRIX analysis, we found that research data repository domains are usually less visible than the domains of their home institutions. At the same time, we know that subdomains inherit some of the visibility of their parent domain. As such we recommend using subdomains of the larger institution to host repositories, for maximum benefit, whenever possible.

2.4 Outlook

Increasing data searchability and findability for the KonsortSWD members remains an important objective of this measure, and we perceive that one of the best ways to achieve this for this community is to do it via a specific strategy that includes a set of guidelines that the KonsortSWD community can follow. The strategy parts we propose allows us to group certain considerations to reach our objective and make them easier - and modular, to some extent - for the community to adopt.

Based on the work in this section, we can confirm that the decentralized approach to SEO is already the status quo in the KonsortSWD community, which is in line with the option that we propose in the strategy (as represented in Table 8). Moreover, since increasing data searchability and findability is not a static, "one and done", but a more dynamic objective. To this effect, we foresee that working with individual community members on concrete actions that suit their cases (providing additional data for their SEO, for example), or more hands-on recommendations about measures they can implement - individually or as a package - are a good way to (dynamically) adjust the data findability, as search and other practices affecting it change.

3 Milestone 3: Enrichment of KonsortSWD metadata schemas

3.1 Introduction

Metadata plays an important role for (digital) resources as it serves a certain purpose and enables a certain functionality (Coyle, 2004). Moreover, metadata and its presentation is another key issue to improve visibility. Metadata standards, on the other hand, build on metadata elements and provide specific functions based on those elements, such as descriptive, administrative, technical, etc. (Higgins, 2007).

Although we are working on improving data searchability and findability on the Web, the topic of metadata is quite central to our measure. As seen in many parts of the report so far, we have dealt with it from different, often complementary perspectives, be it in our survey (explore the important metadata elements, or adopted metadata standards), interviews (the rationale for such standards), or SEO analysis (the type of keywords used during data search). It is, thus, important not only to discuss certain metadata standards, but also to give recommendations for metadata practices that communities could adopt is an important aspect to treat.

As seen in **Section 1.5** on the keyword analysis, metadata allows users on the most common search paths to research data. These cases contain keywords, description, name, and spatial information, and synonyms or acronyms in either the title or the description elements to improve the dataset findability. Milestone 3 is the moment we specifically turn our attention to the role of metadata - be it on the impact that individual metadata elements or metadata standards play in the context of data searchability and findability. As a result, in this section we present three hands-recommendations that correspond to metadata and metadata standards cases, as ways to improve data visibility for the community.

3.2 Web vocabularies: Schema.org and Dublin Core

The way data collections are described via metadata standards is another important aspect to handle in practice. In this context, we are referring to a set of vocabularies (the ones for the web), that are dedicated to represent and provide functionality (thinkg search, retrieval, etc.) to the web pages that describe the data. The integration of web vocabularies, such as Dublin Core and Schema.org, allows the data to be indexed with standard vocabularies, making them directly available to metasearch engines. This directly impacts (increases, in this case) the visibility of the data, but it is also positive for the website rating.

The Dublin Core initiative is used very successfully by SSOAR and may lead to inclusion into Google Scholar, one of the academic search engines with the broadest coverage of scholarly publications (out of 12 such search engines Gusenbauer 2019 compared). On the other hand, Schema.org is a relatively cheap measure to increase data findability. We have already successfully tested its application with da|ra and Gesis Dataset Search. In the latter case, the application of this vocabulary almost doubled the number of users using the service.

At this point, it is important to discuss the effort required to apply these (or any, for that matter) vocabularies. Specifically, determining the number of metadata elements from a vocabulary, and the ones that are both effective and common enough across the domains remain an important task. Often, different communities can provide different sets of metadata elements, and this can be as a result of the metadata practices for those communities, the specifics of certain (domains) datasets, and so on. As a result, we need to identify and recommend a relatively minimal set of impactful metadata elements. For Schema.org this includes elements such as name description, license, provider, creator, date published, and identifier. Moreover, some of the tools we recommend, such as Google Search Console Rich Results Test, enable the verification of a metadata description for a given vocabulary (Schema.org in this case).

3.3 Title and description: What should be in there?

In our Milestone 1 report **on keywords from section 1.5**, one key finding was that users not only search for names and keywords, they also search for acronyms and other types of informal names, including different language versions. As such, it is imperative that the metadata for a given dataset does include proper name, acronyms and other alternative names in a prominent position, ideally the title field. For example, the Big Five Inventory is also known as BFI or BFI-10, as such the Title should be "Big Five Inventory (BFI-10)", as BFI is already a part of BFI-10. The use of parentheses, dashes, or similar symbols is optional and can be freely used to improve readability.

Other than that the Title should be as short as possible. Filler words, such as the, and, of, etc. should be avoided. In direct competition, "The Big Five Inventory of Personality" loses out to the shorter "Big Five Inventory", all other factors being equal. Title and Description have length limitations, which can be checked through the appropriate tools (see section using tools to find and fix). It is imperative to keep within these restrictions, as datasets with incorrect metadata are often omitted from the indexing process. If the alternative titles do not fit into the title line, it is also possible to use the description to list more name variants.

From our analysis it is hard to say what should be in the description, as only relatively few of the queries are matched to description content. This is not to say that the description is not important, but it might indicate that it is more important to list general keywords (see section keywords) as well as constructs used in the dataset. Other than that it should focus on telling human readers what to expect from the data.

3.4 Keywords

We already mentioned the importance of keywords in describing the data on the Web. We had a chance to understand the keywords usage practices of our communities via the interview sessions, and complemented this understanding via a dedicated analysis (based on the search logs from several members) on the keyword description patterns. We foresee keywords as an important part of the hands-on recommendations to the KonsortSWD communities.

In most cases of KonsortSWD, we see the use of websites (and not dedicated data repository solutions) to provide access to the data collections. In any case, there are few considerations, in addition to the "organic" keywords describing a dataset that one can adopt to reach a better visibility of their (dataset) work. The first recommendation is the usage of keywords for the dataset landing pages. Also, for this page, the most important keyword is the name of the dataset, so place this in the URL as well as an H1 header. We have noticed - and it was explicitly mentioned during the interviews - that researchers rely on the dataset name to find as this is what is used in the specific domains, and, having in mind the pattern of searching for data based on a literature item, the name of the dataset as mentioned in the literature/publication plays an important role in identifying the right dataset.

In addition to the dataset keywords and its name, another good step is to explicitly indicate that this is a dataset (and point out the difference to a potential publication with the same name, for example) by using terms such as "dataset", "datensatz", or "datenbank". Moreover, keywords specific to the social sciences could include "question", "questionnaire", "skala", as applicable.

On a final note, keyword optimization is a central point in any SEO strategy. In our study, we found that most content queries for data come via the word "questionnaire". Other keywords, such as "survey", "data", etc., which are also often used in data search, do not lead to the website for the simple reason that they are not mentioned on the landing or detail pages. It is recommended to perform a systematic analysis of relevant keywords and then integrate them into the pages. Other avenues, such as AdWords' free suggestion service and Google Trends, can also be used to this end.

3.5 Concluding remarks

In the context of data searchability and findability, the role of metadata is quite important. Namely, not only the mere adoption of metadata standards, but also specific practices on using metadata elements makes a difference for the data visibility end result. The Dublin Core and Schema.org remain two of the most popular metadata standards for digital resources, and one of the first steps in the right direction towards a data visibility path. Moreover, specific metadata elements from such standards, such as title, description, and keywords, require specific practices, as shown in this section, in order to suit the data visibility objective.

References

1. Baker, M., 2013. Every page is page one. XML Press.
2. Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One*, *9*, e92590.
3. Biribicchi, A. (2020, October 06). How does owning multiple websites for a single company impact SEO?. Amazee Metrics. https://www.amazeemetrics.com/en/blog/seo-for-companies-with-multiple-websites-amazee-metrics/.
4. Chung, C. (2018) Making sense of global integration and local responsiveness in international HRM research. International Journal of Multinational Corporation Strategy, 2 (2). 153. ISSN: 2059-1047. DOI: https://doi.org/10.1504/IJMCS.2018.10010737. Available at http://centaur.reading.ac.uk/72745/.
5. Cojocariu, A. (n.d.). Sitemaps & SEO: Are Sitemaps Still Important for SEO in 2021?. Cognitive SEO. Retrieved December 10, 2021, from https://cognitiveseo.com/blog/22867/seo-sitemaps/.
6. Coyle, K. (2004). Metadata: Data With a Purpose. http://www.kcoyle.net/meta_purpose.html.
7. Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., ... others. (2019). A data citation roadmap for scholarly data repositories. *Scientific data, 6*, 1–9.
8. Friedrich, T. (2020). Looking for data.
9. GoFAIR. (n.d.). FAIR Principles. https://www.go-fair.org/fair-principles/
10. 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.
11. Gusenbauer, M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics 118, 177–214 (2019). https://doi.org/10.1007/s11192-018-2958-5.
12. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems, 17*, 107–145.
13. Higgins, S. (2007). What are Metadata Standards? DCC. https://www.dcc.ac.uk/guidance/briefing-papers/standards-watch-papers/what-are-metadata-standards.
14. Kern, D., & Hienert, D. (2018). Understanding the information needs of social scientists in Germany. *Proceedings of the Association for Information Science and Technology, 55*, 234–243.
15. Kern, D., & Mathiak, B. (2015). Are there any differences in data set retrieval compared to well-known literature retrieval? *International Conference on Theory and Practice of Digital Libraries*, (pp. 197–208).
16. Khan, N., Pink, C. J., & Thelwall, M. (2020). Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations. *Patterns, 1*, 100007.
17. Krämer, T., Papenmeier, A., Carevic, Z., Kern, D., & Mathiak, B. (2021). Data-Seeking Behaviour in the Social Sciences. *International Journal on Digital Libraries, 22*, 175–195.
18. Kratz, J. E., & Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PLoS One, 10*, e0117619.
19. Macdonald, Craig (2011, March 17). Global SEO Programs Decentralize to Regional Governance Model. Search Engine Watch. https://www.searchenginewatch.com/2011/03/17/global-seo-programs-decentralize-to-regional-governance-model/
20. Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository.
21. Onaifo, D. and Rasmussen, D., 2013. Increasing libraries' content findability on the web with search engine optimization. *Library Hi Tech.*

22. Ralambondrainy, H. (1995). A conceptual version of the K-means algorithm. *Pattern Recognition Letters*, 16, 1147-1157. doi:https://doi.org/10.1016/0167-8655(95)00075-R.
23. Sickler, J. (2021, January 8). Top 15 Most Important SEO Metrics to Track Performance. Terakeet.
24. Statista. Worldwide desktop market share of leading search engines from January 2010 to September 2021. https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/
25. What Are Internal Links (Inlinks & Outlinks)? (n.d.). Botify. Retrieved December 10, 2021, from https://www.botify.com/learn/basics/internal-links.
26. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), pp.1-9.

Appendix

Table of Figures

Table of Tables