Dataset Retrieval: Informationsverhalten von Datensuchenden und das Ökosystem von Data-Retrieval-Systemen

Dorothea Strecker

Kurzfassung: Verschiedene Stakeholder fordern eine bessere Verfügbarkeit von Forschungsdaten. Der Erfolg dieser Initiativen hängt wesentlich von einer guten Auffindbarkeit der publizierten Datensätze ab, weshalb Dataset Retrieval an Bedeutung gewinnt. Dataset Retrieval ist eine Sonderform von Information Retrieval, die sich mit dem Auffinden von Datensätzen befasst. Dieser Beitrag fasst aktuelle Forschungsergebnisse über das Informationsverhalten von Datensuchenden zusammen. Anschließend werden beispielhaft zwei Suchdienste verschiedener Ausrichtung vorgestellt und verglichen. Um darzulegen, wie diese Dienste ineinandergreifen, werden inhaltliche Überschneidungen von Datenbeständen genutzt, um den Metadatenaustausch zu analysieren.

Abstract: Various stakeholders are calling for better availability of research data. The success of these initiatives depends largely on good discoverability of published datasets, which is why Dataset Retrieval is gaining in importance. Dataset Retrieval is a special form of Information Retrieval that is concerned with finding datasets. This paper summarizes recent research on the information behavior of data users. Subsequently, two search services with different objectives are presented and compared. In order to show how these services interconnect, overlaps in content are used to analyze metadata exchange between them.

Einleitung

Um die Nachvollziehbarkeit und Nachnutzbarkeit von Forschungsergebnissen zu fördern, fordern verschiedene Stakeholder eine bessere Verfügbarkeit von Forschungsdaten. Der Erfolg dieser Initiativen hängt allerdings wesentlich von einer guten Auffindbarkeit der publizierten Datensätze ab, denn der Kreis von der Publikation bis zur Nachnutzung bestehender Forschungsdaten lässt sich nur schließen, wenn für die jeweilige Fragestellung passende Daten gefunden werden können.

Verschiedene Aspekte von Information Retrieval werden schon lange beforscht, inzwischen hat sich ein eigenständiges Forschungsfeld etabliert (Luk 2022). Im Zuge der wachsenden Anzahl publizierter Forschungsdaten kommen auch Retrieval-Systeme auf, die sich auf das Auffinden von Forschungsdaten spezialisiert haben. Diese Entwicklung bringt neue Fragestellungen und Herausforderungen mit sich.

 ISSN: 1860-7950

Beispielsweise müssen Informationsbedürfnisse und -verhalten von Datensuchenden analysiert werden, um bedarfsgerechte Angebote entwickeln zu können. Außerdem muss sich ein stabiles Ökosystem von Suchdiensten entwickeln, die variierende Bestände durchsuchbar machen, wie es beispielsweise im Bibliotheksbereich schon lange existiert.

Fragestellungen

In diesem Beitrag soll darum dargestellt werden, was aktuell über das Informationsverhalten von Datensuchenden bekannt ist. Anschließend werden beispielhaft zwei Suchdienste – PANGAEA und Google Dataset Search – mit besonderem Fokus auf Unterschiede in ihrer Datengrundlage und Funktionsweise vorgestellt. Um zu zeigen, wie diese Dienste ineinandergreifen, werden inhaltliche Überschneidungen von Datenbeständen genutzt, um den Metadatenaustausch zu analysieren.

Literaturübersicht

Information Retrieval und Dataset Retrieval

Allgemein bezeichnet Information Retrieval (IR) den Prozess, bei dem gezielt nach Informationsobjekten in einer Sammlung gesucht wird, die ein bestimmtes Informationsbedürfnis abdecken (Meadow et al. 2007). Systeme, die IR ermöglichen, repräsentieren und organisieren Informationen auf unterschiedliche Weise, um sie durchsuchbar zu machen, und bieten verschiedene Einstiege für die Suche an. Die Ursprünge von IR sind eng mit der Entwicklung des Bibliothekswesens verknüpft (Sanderson und Croft 2012). Im Laufe der Zeit haben sich IR-Systeme wesentlich weiterentwickelt, und IR hat sich als eigenständige Forschungsdisziplin etabliert (Luk 2022).

Dataset Retrieval (DR) kann als Sonderform von IR betrachtet werden, die sich auf das Auffinden von Informationsobjekten eines besonderen Typs (Datensätze) spezialisiert (Chen et al. 2019; Kunze und Auer 2013). DR ist ein vergleichsweise neues Forschungsfeld, das sich in der Anfangszeit vorwiegend mit technischen Herausforderungen beschäftigt hat; soziale Aspekte wie Informationsbedürfnisse und -verhalten wurden erst in den letzten Jahren systematisch beforscht (Gregory, Cousijn, et al. 2020). Diese Aspekte sind jedoch wichtig, um das Informationsverhalten von Datensuchenden zu verstehen und daraus Anforderungen an DR-Systeme abzuleiten.

Funktionsweise von DR-Systemen

Grundsätzlich funktionieren DR-Systeme wie andere IR-Systeme: Eine Anfrage durch Nutzende erfolgt in den Phasen *querying* (Stellen der Anfrage durch den*die Nutzer*in), *query handling* (Bearbeitung der Anfrage durch den Suchdienst), *data handling* (Identifizieren von Treffern durch den Suchdienst) und *results presentation* (Anzeigen der Treffer durch den Suchdienst) (Chapman et al. 2020). Das DR-System erstellt im Voraus einen Index der durchsuchbaren Dokumente und beinhaltet Komponenten, die Anfragen mit dem Index abgleichen und Trefferlisten erstellen

 ISSN: 1860-7950

(Chen et al. 2019). Der Index kann auf verschiedene Weise erzeugt werden, beispielsweise durch das gezielte Harvesten von Metadaten über die Schnittstellen ausgewählter Dienste oder durch Crawls, die regelmäßig eine große Zahl von Repositorien ansteuern (Chapman et al. 2020). DR basiert aktuell überwiegend auf strukturierten Metadaten, die Repositorien bereitstellen (Chapman et al. 2020; Devaraju und Berkovsky 2018). Ansätze zur Anreicherung des Indexes wie die Volltextindexierung von Publikationen, die auf den durchsuchbaren Datensätzen aufbauen, sind bisher noch wenig verbreitet (Khalsa, Cotroneo, und Wu 2018).

Aktuell sind die Suchfunktionalitäten von DR-Systemen vorwiegend auf Keywordsuche und facettierte Navigation beschränkt (Devaraju und Berkovsky 2018). Um den Ansprüchen von Nutzenden gerecht zu werden, sollten Repositorien möglichst mehrere Sucheinstiege anbieten, beispielsweise neben einem einfachen Suchschlitz auch eine erweiterte Suche, die Suche über eine Karte oder Browsing nach Facetten (Wu et al. 2019) – was viele Repositorien bereits umgesetzt haben (Khalsa, Cotroneo, und Wu 2018).

Evaluation von DR-Systemen

Ein ungelöstes Problem ist die Evaluation von DR-Systemen. In einer Umfrage von 2018 unter 98 Repositorienbetreiber*innen gab nur etwa ein Drittel der Befragten an, das DR-System in der Vergangenheit evaluiert zu haben, und nur etwa die Hälfte war sich sicher, dass die meisten Anfragen durch das System für Nutzer*innen zufriedenstellend beantwortet werden (Khalsa, Cotroneo, und Wu 2018). Ein Grund dafür könnte sein, dass verbreitete Metriken und Benchmarks für die Evaluation von IR-Systemen häufig auf Textdokumente ausgelegt sind und nicht immer direkt auf Datensätze übertragbar sind (Chapman et al. 2020).

Trotz dieser Herausforderungen gibt es einige Publikationen, die Aspekte von DR-Systemen evaluieren. So zeigte sich beispielsweise, dass die Relevanzbewertung von Treffern noch besser an spezifische Besonderheiten von Datensätzen angepasst werden könnte (Dede Şener, Ogul, und Basak 2022). Auch verschiedene Ansätze für Recommender-Services, die ähnliche Datensätze aufzeigen, wurden erprobt (Wang, Huang, und Harmelen 2020).

Es gibt erste Hinweise darauf, dass integrierte Retrieval-Systeme, die sowohl Daten- und Textpublikationen auffindbar machen, unterschiedlich gut mit diesen Dokumenttypen umgehen. In einem solchen integrierten System waren einige populäre Datensätze sehr gut, viele andere dagegen weniger gut auffindbar (Roy, Carevic, und Mayr 2022). Die Auffindbarkeit bei Textpublikationen war im Vergleich gleichmäßiger verteilt, außerdem führte die Suche nach Textpublikationen wesentlich häufiger zu direkten Interaktionen mit Inhalten.

Informationsverhalten im Kontext von Forschungsdaten

Borst und Limani (2020) stellen *data search* als dreiteiliges Konzept dar, das die Aspekte *discovery* (Suche nach Datensätzen in einem Retrievalsystem), *exploration* (Inhalt und Struktur eines Datensatzes mithilfe der Metadaten erkunden) und *analysis* (Teile eines Datensatzes gemäß einer Fragestellung und datensatzspezifischer Eigenschaften auswählen) umfasst. In diesem Beitrag wird vorwiegend der Aspekt *discovery* behandelt.

 ISSN: 1860-7950

Forschende suchen Daten für diverse Nachnutzungsszenarien (Gregory et al. 2019). Diese lassen sich grob in *background*-Nutzung (wie etwa die Kalibrierung von Messinstrumenten oder den Einsatz von Daten in der Lehre) und *foreground*-Nutzung (wie etwa das Beantworten eigener, neuer Fragestellungen anhand nachgenutzter Daten) unterteilen (Gregory, Cousijn, et al. 2020).

Aktuell basieren DR-Ansätze überwiegend auf Erfahrungen mit der Suche nach Textdokumenten (Borst und Limani 2020). Die Suche nach Daten unterscheidet sich jedoch in einigen zentralen Aspekten von der Suche nach Textpublikationen. So sind Daten im Vergleich komplexere Objekte, da sie auch Begleitmaterial wie Codebücher beinhalten können (Carevic, Roy, und Mayr 2020). Zudem sind die Anforderungen an DR deutlich vielfältiger: DR-Systeme werden je nach Informationsbedürfnis auch mit Aspekten wie Provenienz, Qualität, Granularität und Interoperabilität von Daten konfrontiert (Chapman et al. 2020). Darum werden Metadaten bei DR wichtiger eingeschätzt, da Forschende über Metadaten den Kontext der Daten verstehen und die Nutzbarkeit bewerten können (Kern und Mathiak 2015).

Suchstrategien

In einer Umfrage von 2020 unter 1.677 Forschenden berichtete die Mehrzahl der Befragten, dass sie die Suche nach Daten herausfordernd oder sogar schwierig finden; ein Drittel gab als Grund dafür unzureichende Suchdienste an (Gregory, Groth, et al. 2020). Forschende nutzen für das Auffinden von Datensätzen verschiedene Quellen und Strategien. Häufig genutzt werden beispielsweise persönliche Kontakte, Verweise in Textpublikationen oder Websuchmaschinen (Friedrich 2020; Gregory, Groth, et al. 2020; Krämer et al. 2021). Die Nutzung spezialisierter Dienste wie Forschungsdatenrepositorien oder Suchmaschinen für Forschungsdaten ist demgegenüber weniger verbreitet (Gregory, Groth, et al. 2020). Insbesondere zu Beginn eines Suchprozesses nutzen Forschende häufig Websuchmaschinen, sowohl um Daten- oder Textpublikationen als auch um spezialisierte Angebote für die Datensuche zu finden (Krämer et al. 2021). Allgemeine Berufserfahrung scheint einen positiven Einfluss auf die Diversität der Quellen zu haben, die Forschende für DR nutzen (Friedrich 2020). Eine erste Metaanalyse zeigt, dass sich die Informationsbedürfnisse und Herausforderungen von Forschenden und Forschungsdatenexpert*innen wie Data Librarians, ähneln (Sun et al. 2022). Allerdings nutzen sie unterschiedliche Strategien, um geeignete Daten zu finden – so nutzen Data Librarians weniger häufig Websuchmaschinen.

Wenn Forschende Websuchmaschinen für DR nutzen, ergänzen sie die Suchanfrage häufig um Begriffe wie ‚data' oder ‚dataset', um ihr Bedürfnis zu präzisieren (Koesten et al. 2017; Krämer et al. 2021). Websuchmaschinen spielen auch eine wichtige Rolle beim Auffinden spezialisierter Datensuchdienste. Die Logfiles von Open-Data-Portalen zeigen beispielsweise, dass die meisten Nutzer*innen über Websuchmaschinen auf die Seiten der Portale gelangten (Kacprzak et al. 2019; Koesten et al. 2017). Obwohl viele Nutzer*innen die Open-Data-Portale und ihre Inhalte bereits kannten, was die gehäufte Nennung dieser Portale in den Anfragen an Websuchmaschinen nahelegt, zogen sie offenbar die Suchfunktion der Websuchmaschinen den Open-Data-Portalen vor. Nutzer*innen, die von externen Diensten wie Websuchmaschinen auf Inhalte von Open-Data-Portalen verwiesen wurden, brachen ihre Suche allerdings auch schneller erfolglos ab als Nutzer*innen, die den gesamten Suchprozess im Portal durchführten (Ibáñez und Simperl 2022).

 ISSN: 1860-7950

Im Vergleich zu Suchanfragen an Websuchmaschinen sind Suchanfragen an spezialisierte Datensuchdienste in der Tendenz kürzer. Suchanfragen an Open-Data-Portale umfassten im Durchschnitt beispielsweise zwischen 1,63 und 2,52 Wörter (Kacprzak et al. 2017). Forschende scheinen kurze Suchanfragen bewusst einzusetzen, um lange Trefferlisten zu erstellen, die sie dann manuell auf Nutzbarkeit überprüfen (Kacprzak et al. 2019; Koesten et al. 2017). Beim DR über spezialisierte Datensuchdienste enthalten Suchanfragen außerdem häufiger Ziffern, räumliche oder zeitliche Einschränkungen in verschiedenen Granularitätsstufen oder die Nennung spezieller Datentypen oder -formate (Carevic, Roy, und Mayr 2020; Kacprzak et al. 2019). Suchanfragen an Datensuchdienste scheinen insgesamt ein breiteres Spektrum von Themen abzudecken (Kacprzak et al. 2019). Bei der Suche nach Daten sind sich Suchanfragen, die innerhalb einer Session gestellt werden, thematisch ähnlicher als bei der Suche nach Textpublikationen (Carevic, Roy, und Mayr 2020). Interaktionspfade bei der Suche nach Daten- und Textpublikationen ähneln sich generell, allerdings werden Anfragen bei der Suche nach Textpublikationen häufiger umformuliert und neu gestellt (Carevic, Roy, und Mayr 2020).

Unklar ist noch, in welchem Umfang *known-item search* auftritt – die Suche nach einem bestimmten, bereits bekannten Datensatz. Dass beim DR explorative Ansätze verfolgt werden, zeigt sich beispielsweise daran, dass Suchende durch kurze Suchanfragen bewusst lange Trefferlisten erzeugen (Kacprzak et al. 2017, 2019). Jedoch gibt es keine Einigkeit über die Verbreitung von *known-item search*: Analysen zeigen, dass Anfragen an spezialisierte Datensuchdienste selten angepasst und verändert neu gestellt werden (Koesten et al. 2017; Carevic, Roy, und Mayr 2020). In diesem Bereich ist mehr Forschung erforderlich, um *known-item search* in DR-Systemen besser unterstützen zu können.

Die hier beschriebenen Studien des Informationsverhaltens in Bezug auf DR legen nahe, dass es bisher kaum einheitliche und bewährte Strategien für die Suche nach Datensätzen gibt (Krämer et al. 2021).

Arten von DR-Systemen

Grundlegend gibt es verschiedene Arten von DR-Systemen. Sie bilden ein komplexes Ökosystem, in dem jeder Dienst abhängig von der Mission und Nutzer*innengruppe verschiedene Bedürfnisse abdeckt. Beispielsweise können DR-Systeme in zentrale und dezentrale Angebote unterteilt werden (Chapman et al. 2020). Sie divergieren darin, ob ein zentraler Bestand durchsuchbar gemacht wird oder ob Inhalte mehrerer Quellen zusammengeführt werden (Borst und Limani 2020). Suchdienste unterscheiden sich außerdem in ihrem disziplinären Fokus: Disziplinübergreifende Angebote machen Bestände mehrerer Disziplinen durchsuchbar, während sich disziplinspezifische DR-Systeme auf Daten einer Disziplin konzentrieren.
Im Folgenden werden beispielhaft zwei DR-Systeme vorgestellt und verglichen, die verschiedenen Ansätzen folgen. Google Dataset Search, ein sehr umfassender dezentraler und disziplinübergreigender Dienst, wird PANGAEA, einem zentralen und disziplinspezifischen Dienst gegenübergestellt.

 ISSN: 1860-7950

Dezentral und disziplinübergreifend: Google Dataset Search

2018 wurde Google Dataset Search in der Beta-Version veröffentlicht. Der Dienst nahm 2020 offiziell den Betrieb auf (Noy und Benjelloun 2020).

Datenquellen: Der Index von Google Dataset Search basiert zwar auf Crawls, im Gegensatz zur Google Websuchmaschine aber zugleich auf strukturierten Metadaten. Datenlieferanten zeichnen Metadaten nach bestimmten Standards (schema.org und DCAT) aus und stellen diese strukturierten Metadaten über die landing pages individueller Datensätze zur Verfügung. Ein Crawler sammelt diese Metadaten, die anschließend verarbeitet und indexiert werden. Anfragen von Nutzenden werden mit dem Index abgeglichen und Ergebnisse nach Relevanz sortiert angezeigt. Die Relevanzbewertung basiert auf demselben Ansatz wie die Google Websuchmaschine, bezieht jedoch auch Aspekte von Metadatenqualität ein (Brickley, Burgess, und Noy 2019).
Die Größenverteilung der Datenquellen in Google Dataset Search ist stark verzerrt: Die 20 größten Datenlieferanten machen 78 % des gesamten Bestands aus (Benjelloun, Chen, und Noy 2020). Welche Inhalte der Google Dataset Search Index im Detail umfasst, ist unbekannt. Google Dataset Search indexiert auch Aggregatoren von Metadaten, beispielsweise DataCite. Das führt dazu, dass die Registrierung von DOIs neben anderen Vorteilen auch als Strategie empfohlen wird, um Datensätze über Google Dataset Search auffindbar zu machen (Masson et al. 2021).

Metadaten: Da der Dienst disziplinübergreifend konzipiert ist, ist das verwendete Metadatenschema sehr generisch. Außerdem gibt es nur zwei Pflichtfelder: Titel und Beschreibung. Daraus ergeben sich Probleme in Bezug auf die Metadatenqualität, insbesondere die Vollständigkeit. So sind beispielsweise Lizenzinformationen nur für 34 % der Datensätze verfügbar (Benjelloun, Chen, und Noy 2020).

Sucheinstiege: Der Sucheinstieg bei Google Dataset Search ist auf eine einfache Keywordsuche beschränkt, auch die Präsentation der Ergebnisse ist einfach gehalten. Eine niedrigschwellige Anwendung muss jedoch kein Nachteil sein, sondern könnte die Nachnutzung von Daten popularisieren und Anreize für das Publizieren von Datensätzen schaffen (Canino 2019).

Ergebnispräsentation: Ergebnisse werden in Google Dataset Search mit den vorhandenen Metadaten dargestellt (siehe Abbildung 1). Sofern vorhanden, wird die räumliche Abdeckung des Datensatzes auf einer Karte gezeigt.

Zentral und disziplinspezifisch: PANGAEA

PANGAEA nahm 1994 den Betrieb auf. Das disziplinspezifische Forschungsdatenrepositorium hat sich auf das Publizieren georeferenzierter Datensätze aus Bereichen der Erdsystemwissenschaft spezialisiert (Diepenbroek et al. 2002).

Datenquellen: Der Suchdienst umfasst die in PANGAEA publizierten Datensätze.[1]

Metadaten: Metadaten in PANGAEA sind an die spezifischen Besonderheiten der publizierten Datensätze angepasst. So bietet das Format, das zur Beschreibung von Datensätzen genutzt wird, die Möglichkeit, detaillierte Informationen zu dem dokumentierten Ereignis, Expeditionen und den Erhebungsmethoden anzugeben (Diepenbroek et al. 2017). Umfassende Kurations-

[1] PANGAEA Terms of Use: https://www.pangaea.de/about/terms.php

 ISSN: 1860-7950

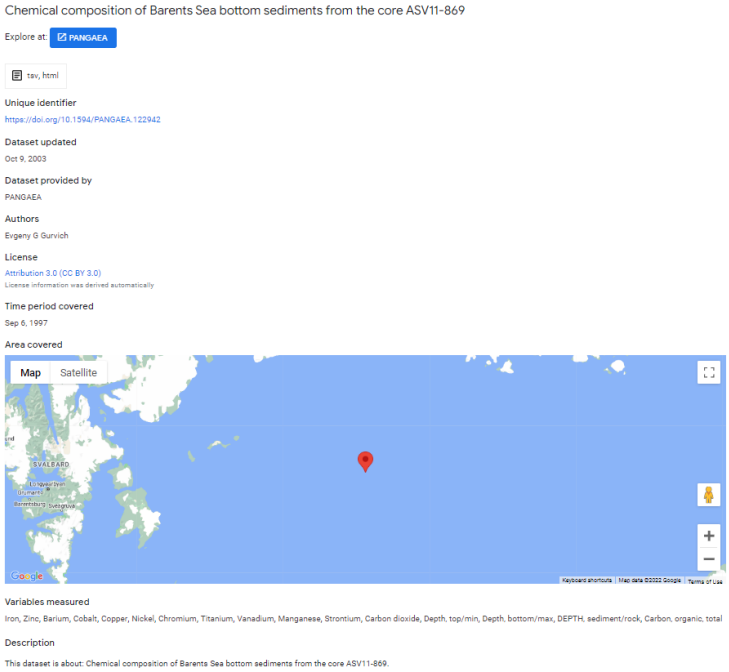

Abbildung 1: Ergebnispräsentation für den Datensatz https://doi.org/10.1594/PANGAEA.122942 in Google Dataset Search

und Qualitätssicherungsprozesse stellen die Nützlichkeit der Metadaten sicher.
PANGAEA macht Metadaten offen (unter einer CC0-Lizenz) über Schnittstellen[2] zugänglich. Der Dienst legt Wert auf die Interoperabilität der Metadaten und liefert sie in entsprechenden Formaten an verschiedene Aggregatoren aus, beispielsweise GBIF oder OpenAIRE. Metadaten werden außerdem nach schema.org ausgezeichnet über die landing pages bereitgestellt. Diese Aktivitäten führen dazu, dass die publizierten Datensätze auch in anderen dezentralen DR-Systemen gut auffindbar sind.

Sucheinstiege: PANGAEA bietet neben einer einfachen Keywordsuche auch Browsing nach Disziplinen und über eine Karte an. Das PANGAEA Data Warehouse ermöglicht Datennutzenden außerdem das effiziente Zusammenstellen von Datensätzen nach selbst definierten Parametern wie beispielsweise alle Messungen einer bestimmten Größe für eine Region.[3]

Ergebnispräsentation: PANGAEA stellt den Kontext von Datensätzen beispielsweise durch Details zur Datenerhebung oder Projekte und andere Publikationen, die mit dem Datensatz in Zusammenhang stehen (siehe Abbildung 2), sehr ausführlich dar. Die räumliche Abdeckung des Datensatzes wird auf einer Karte abgebildet. Außerdem werden Zitationsempfehlung, Ansichts- und Downloadzahlen sowie eine Übersicht der im Datensatz enthaltenen Variablen angezeigt.

[2]Es werden verschiedene Schnittstellen angeboten, siehe die API-Dokumentation: https://ws.pangaea.de/
[3]PANGAEA Data Warehouse: https://www.pangaea.de/tools/

 ISSN: 1860-7950

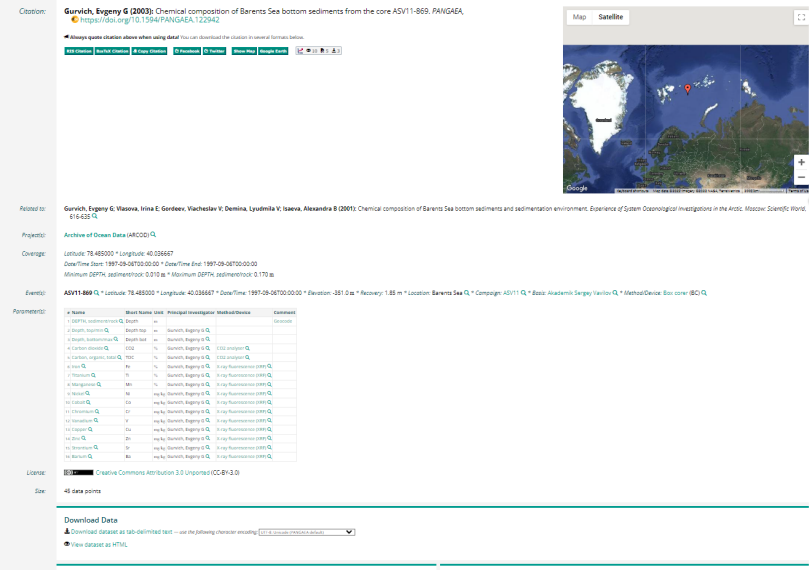

Abbildung 2: Ergebnispräsentation für den Datensatz https://doi.org/10.1594/PANGAEA.122942 in PANGAEA

Methode

Zwischen den hier vorgestellten Diensten gibt es inhaltliche Überschneidungen bei den Datenbeständen: Durch das Bereitstellen von Metadaten über die landing pages sorgt PANGAEA dafür, dass Datensätze auch in Google Dataset Search indexiert sind. Im Folgenden wird näherungsweise untersucht, wie erfolgreich Metadaten an Google Dataset Search übergeben werden. Zu diesem Zweck werden Metadatensätze verglichen, die in beiden Diensten vorliegen.

Datenerhebung

Metadaten für Google Dataset Search stammen aus einem Subset der indexierten Datensätze, das 2020 veröffentlicht wurde (Google Dataset Search 2020). Das Subset umfasst 17 Metadatenelemente für Datensätze, für die eine DOI oder ein anderer Identifier vorlag (3.602.027 Datensätze, Stand 16.10.2020). Da PANGAEA DOIs für alle publizierten Datensätze vergibt, sind diese auch in dem Datensatz enthalten. Zunächst wurden DOIs identifiziert, die in beiden Datenquellen auftreten (DOIs, die am 30.10.2022 über die PANGAEA OAI-OMH-Schnittstelle abrufbar waren sowie DOIs in dem veröffentlichten Subset von Google Dataset Search). Diese Bedingung trifft auf 364.255 Metadatensätze zu.

Metadaten für 10.000 zufällig ausgewählte PANGAEA-Datensätze wurden am 01.11.2022 über die OAI-PMH-Schnittstelle im Format *PANGAEA MetaData* abgefragt.[4]

[4]PANGAEA MetaData: http://ws.pangaea.de/schemas/pangaea/MetaData.xsd

 ISSN: 1860-7950

Analyse

Zunächst wurde die Nutzung der verfügbaren Metadatenelemente im Subset von Google Dataset Search analysiert.

Anschließend wurde beispielhaft an zwei Metadatenelementen untersucht, welche Informationen die beiden vorgestellten Suchdienste für ein zufällig gewähltes Sample von 10.000 Metadatensätzen vorhalten:

- Informationen über die Methode, die der Datenerhebung zugrunde liegt (*method* in PANGAEA und *measurementTechnique* in Google Dataset Search)
- Informationen über Fördermittel, die für die Datenerhebung bereitgestellt wurden (*funder* in PANGAEA und *funder* in Google Dataset Search)

Die zwei beschriebenen Metadatenelemente wurden gewählt, da sie für PANGAEA-Datensätze häufig nicht in der Google Dataset Search vorlagen. Die Auswahl zielt auf eine Analyse möglicher Probleme beim Metadatenaustausch ab.

Ergebnisse

Metadaten in Google Dataset Search

Abbildung 3 zeigt, wie häufig die verfügbaren Metadatenelemente zur Beschreibung der PANGAEA-Datensätze in Google Dataset Search genutzt wurden. Tabelle 1 im Anhang beschreibt die Metadatenelemente und listet ihre Nutzung im Detail auf.

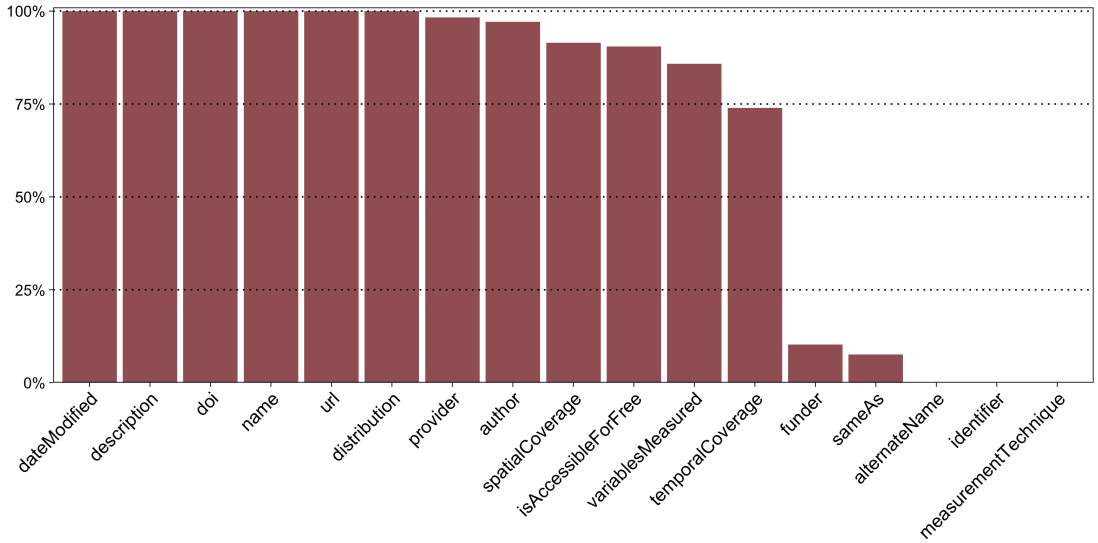

Abbildung 3: Nutzung von Metadatenelementen zur Beschreibung von PANGAEA-Datensätzen in einem Subset von Google Dataset Search (Stand 16.10.2020)

 ISSN: 1860-7950

Fünf der in Abbildung 3 dargestellten Metadatenelemente sind in allen untersuchten Metadatensätzen vorhanden. Dazu zählen neben den zwei Pflichtelementen *name* und *description* auch *dateModified*, *doi* und *url*. Fünf weitere Elemente werden in über 90 % der Metadatensätze genutzt: *distribution*, *provider*, *author*, *spatialCoverage* und *isAccessibleForFree*. Die Elemente *variablesMeasured* und *temporalCoverage* liegen für die Mehrzahl der Metadatensätze vor.

Im Vergleich deutlich weniger genutzt werden *funder* und *sameAs*, während drei Metadatenelemente in diesem Bestand nicht vorkommen (*alternateName*, *identifier*, *measurementTechnique*). Die mangelnde Nutzung einiger dieser Elemente kann damit erklärt werden, dass sie nicht auf alle Datensätze zutreffen – so hat nicht jeder Datensatz alternative Titel, URLs oder Identifier, die in den Metadaten abgebildet werden können.

Vergleich zwischen PANGAEA und Google Dataset Search

Abbildung 4 zeigt anhand eines zufällig gewählten Samples von 10.000 Datensätzen, die in beiden Datenquellen beschrieben sind, Unterschiede in der Verfügbarkeit von Informationen zu Förderung (*funder*) und Methoden (*method*).

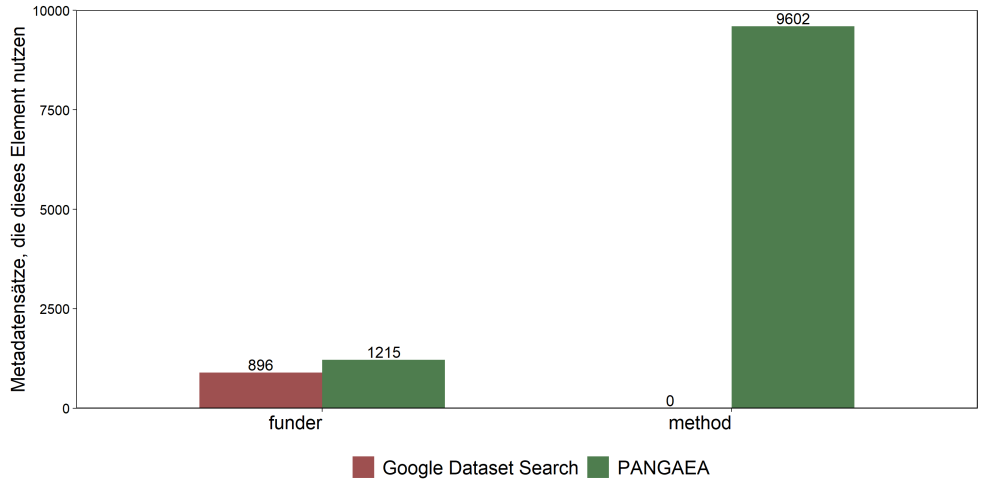

Abbildung 4: Informationen zu Förderern und Methoden in Google Dataset Search und PANGAEA (Sample von 10.000 Metadatensätzen)

Informationen zu Fördermitteln sind in PANGAEA etwas häufiger vorhanden als in Google Dateset Search; aber auch in PANGAEA sind diese Angaben nicht weit verbreitet. Unterschiede zwischen den beiden DR-Systemen zeigen sich besonders deutlich in der Verfügbarkeit von Informationen zu Methoden, die der Datenerhebung zugrunde liegen. Während diese Angaben in PANGAEA für fast alle Datensätze vorliegen (96,02 % ; n = 9.602), fehlen sie in Google Dataset Search komplett.

 ISSN: 1860-7950

Diskussion

DR ist ein schnell wachsendes Forschungsfeld innerhalb von IR, was sicher von der steigenden Anzahl verfügbarer Datensätze durch Open-Science-Initiativen beeinflusst wird.
DR-Systeme orientieren sich bisher stark an IR-Systemen, werden jedoch nicht immer den objekttypspezifischen Anforderungen gerecht, die Datensätze mit sich bringen können. Zu den aktuellen Herausforderungen im Bereich DR zählt das Fehlen von Suchansätzen, die über Keywordsuche und Filteroptionen hinausgehen, sowie der Fokus auf strukturierte Metadaten, da Metadaten für Forschungsdaten nicht immer in ausreichendem Umfang vorliegen oder Qualitätsanforderungen beziehungsweise Informationsbedürfnissen entsprechen. Auch stellt der Mangel an spezialisierten Evaluationskonzepten für DR-Systeme ein Problem dar. Nur wenige Repositorienbetreiber*innen haben ihr DR-System in der Vergangenheit evaluiert. Dadurch fehlen beispielsweise gesicherte Erkenntnisse über die Eignung des DR-Systems für das erfolgreiche Auffinden geeigneter Datensätze, und damit auch Ansätze für die Verbesserung des Dienstes.

Die Suche nach Daten unterscheidet sich in einigen Aspekten von der Suche nach Textpublikationen. Darum ist es wichtig, das Informationsverhalten von Datensuchenden gezielt zu betrachten. In den vergangenen Jahren gab es eine Reihe von Vorhaben mit dieser Zielsetzung, vor allem in den Sozialwissenschaften. Die Ergebnisse zeigen, dass Datensuchende auf verschiedene Quellen zurückgreifen, auch in Kombination. Dabei nutzen sie nicht nur spezialisierte Datensuchdienste, sondern auch persönliche Kontakte oder Verweise in Textpublikationen. Auch Websuchmaschinen werden bei der Datensuche häufig genutzt. Bei der Nutzung von Websuchmaschinen fallen Muster wie das Ergänzen der Suchanfrage um Begriffe wie ‚data' oder die Namen bekannter Datenrepositorien auf. Suchanfragen in spezialisierten Datensuchdiensten beinhalten häufig Einschränkungen, die sich auf bestimmte Eigenschaften von Datensätzen beziehen, beispielsweise die räumliche oder zeitliche Abdeckung. Eine verbreitete Strategie scheint das bewusste Formulieren kurzer Anfragen zu sein, um lange Trefferlisten zu erzeugen, die dann manuell geprüft werden. Insgesamt sollten spezialisierte Datensuchdienste stärker beworben werden, damit Forschende wissen, wo sie geeignete Daten finden können.

Wie für die Suche nach Textpublikationen haben sich verschiedene Arten von Suchdiensten herausgebildet, die sich beispielsweise im Umfang der indexierten Datenquellen (zentral – dezentral) und dem disziplinären Fokus (disziplinspezifisch – disziplinübergreifend) unterscheiden. Disziplinspezifische und zentrale DR-Systeme wie PANGAEA können stärker als disziplinübergreifende oder dezentrale Dienste (zum Beispiel Google Dataset Search) auf spezielle Informationsbedürfnisse ihrer Nutzenden eingehen, was sich beispielsweise an den durchsuchbaren Metadaten, den zur Verfügung gestellten Sucheinstiegen und der Präsentation der Ergebnisse zeigt. Der Vorteil von dezentralen Diensten ist, dass sie mehrere Datenbestände aggregieren und durchsuchbar machen. Sie können zwar nicht auf spezielle Informationsbedürfnisse eingehen, schaffen aber schwellenarme und bestandsübergreifende Angebote für DR. Dadurch tragen sie zu einer besseren Sichtbarkeit von Datensätzen bei, was auch im Interesse von Repositorien wie PANGAEA ist.

Metadaten der Datensätze, die in PANGAEA publiziert wurden, werden auch von Google Dataset Search indexiert. Der Metadatenaustausch zwischen den Diensten ist sehr effektiv, denn der Beschreibungsgrad von PANGAEA-Datensätzen in Google Datasest Search ist verglichen mit anderen Datenquellen überdurchschnittlich gut. So sind Informationen über Autor*innen nur

 ISSN: 1860-7950

für 14,2 % aller Datensätze im untersuchten Subset von Google Dataset Search vorhanden (Benjelloun, Chen und Noy 2020), während die Angabe für mehr als 90 % der PANGAEA-Datensätze vorliegen. Die Analyse zeigte jedoch, dass Informationen zu Erhebungsmethoden nicht in vollständigem Umfang weitergegeben werden. Eine mögliche Erklärung dafür könnte sein, dass PANGAEA zu dem Zeitpunkt, an dem das Subset erstellt wurde (2020), das Mapping zwischen dem Metadatenschemata PANGAEA MetaData und schema.org noch nicht optimiert hatte. Dadurch könnten Informationen zur angewendeten Methode, die intern vorhanden sind, nicht in die landing pages eingebettet und darum nicht von Crawls erfasst werden.

Metadatenelemente, die in Google Dataset Search genutzt werden, sind überwiegend beschreibend und nicht sehr umfangreich. Sie dienen der Auffindbarkeit, indem identifizierende Eigenschaften (zum Beispiel Titel, Autor*in) sowie der Kontext (zum Beispiel räumliche und zeitliche Abdeckung) abgebildet werden. Demgegenüber können Forschungsdaten in PANGAEA durch ein detailliertes Metadatenschema deutlich präziser beschrieben werden. Darüber hinaus werden bei PANGAEA kontrollierte Vokabulare eingesetzt, die semantische Bezüge zwischen Datensätzen herstellen und die Interoperabilität verbessern.

Fazit

Durch die zunehmende Verfügbarkeit von publizierten Forschungsdaten gewinnt Dataset Retrieval an Bedeutung. Aktuell findet das Thema viel Aufmerksamkeit in der Forschung, zum Beispiel mit Blick auf die Anpassung von Suchdiensten an objekttypspezifische Besonderheiten von Forschungsdaten oder das Informationsverhalten von Datensuchenden. Auch wenn der Metadatenaustausch nicht immer reibungslos abläuft, entwickelt sich nach und nach ein Ökosystem ineinandergreifender DR-Systeme. Einige sind stark an die Bedürfnisse von Datensuchenden angepasst, während andere mehrere Datenbestände aggregieren und schwellenarme Sucheinstiege anbieten. DR-Systeme vereinfachen das Auffinden von Forschungsdaten für die Nachnutzung und tragen so zum Erfolg von Open-Science-Initiativen bei.

Limitationen

Der Vergleich der hier vorgestellten Dienste basiert unter anderem auf einem Subset, das Google Dataset Search 2020 veröffentlicht und seither nicht aktualisiert hat. Metadaten von PANGAEA wurden demgegenüber 2022 gesammelt. Aufgrund dieser zeitlichen Differenz muss der Vergleich als ungefähre Annäherung gelten, die verdeutlichen soll, dass der Metadatenaustausch zwischen Diensten nicht immer reibungslos abläuft.

 ISSN: 1860-7950

Anhang A

Tabelle 1: Beschreibung der Metadatenelemente in Google Dataset Search und deren Vorkommen für PANGAEA-Datensätze

Element	**Beschreibung**	**Verwendung (total)**	**Verwendung (anteilig)**
dateModified	Zeitpunkt, zu dem der Datensatz zuletzt bearbeitet wurde	36.4255	100 %
description	Beschreibung des Datensatzes	36.4255	100 %
doi	DOI des Datensatzes	36.4255	100 %
name	Titel des Datensatzes	36.4255	100 %
url	URL des Datensatzes	36.4255	100 %
distribution	Angaben zum Download des Datensatzes	36.4221	99,99 %
provider	Organisation, die den Datensatz zur Verfügung stellt	358.051	98,3 %
author	Autor*innen des Datensatzes	353.825	97,14 %
spatialCoverage	Räumliche Abdeckung des Datensatzes	333.179	91,47 %
isAccessibleForFree	Indikator, der angibt, ob der Datensatz frei verfügbar ist	329.631	90,49 %
variablesMeasured	Gemessene Variablen, die im Datensatz repräsentiert werden	312.664	85,84 %
temporalCoverage	Zeitliche Abdeckung des Datensatzes	269.335	73,49 %
funder	Fördermittel, die zur Erhebung des Datensatzes zur Verfügung gestellt wurden	37.269	10,23 %
sameAs	Alternative URL des Datensatzes	27.666	7,6 %
alternateName	Alternativer Titel des Datensatzes	0	0 %
identifier	Identifier (außer DOI) des Datensatzes	0	0 %
measurementTechnique	Methode, die zur Messung der repräsentierten Variablen angewendet wurde	0	0 %

 ISSN: 1860-7950

Literaturverzeichnis

Benjelloun, Omar, Shiyu Chen, und Natasha Noy. 2020. „Google Dataset Search by the Numbers.“ In *The Semantic Web – ISWC 2020*, edited by Jeff Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, und Lalana Kagal, 667–82. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_41.

Borst, Timo, und Fidan Limani. 2020. „Patterns for Searching Data on the Web Across Different Research Communities.“ *LIBER Quarterly* 30 (1): 1–21. https://doi.org/10.18352/lq.10317.

Brickley, Dan, Matthew Burgess, und Natasha Noy. 2019. „Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem.“ In *The World Wide Web Conference*, 1365–75. New York: Association for Computing Machinery. https://doi.org/10.1145/3308558.3313685.

Canino, Adrienne. 2019. „Deconstructing Google Dataset Search.“ *Public Services Quarterly* 15 (3): 248–55. https://doi.org/10.1080/15228959.2019.1621793.

Carevic, Zeljko, Dwaipayan Roy, und Philipp Mayr. 2020. „Characteristics of Dataset Retrieval Sessions: Experiences from a Real-Life Digital Library.“ In *Digital Libraries for Open Knowledge*, edited by Mark Hall, Tanja Merčun, Thomas Risse, und Fabien Duchateau, 185–93. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-030-54956-5_14.

Chapman, Adriane, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, und Paul Groth. 2020. „Dataset Search: A Survey.“ *The VLDB Journal* 29 (1): 251–72. https://doi.org/10.1007/s00778-019-00564-x.

Chen, Jinchi, Xiaxia Wang, Gong Cheng, Evgeny Kharlamov, und Yuzhong Qu. 2019. „Towards More Usable Dataset Search: From Query Characterization to Snippet Generation.“ In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2445–48. New York: Association for Computing Machinery. https://doi.org/10.1145/3357384.3358096.

Dede Şener, Duygu, Hasan Ogul, und Selen Basak. 2022. „Text-Based Experiment Retrieval in Genomic Databases.“ *Journal of Information Science*, 01655515221118670. https://doi.org/10.1177/01655515221118670.

Devaraju, Anusuriya, und Shlomo Berkovsky. 2018. „A Hybrid Recommendation Approach for Open Research Datasets.“ In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 207–11. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3209219.3209250.

Diepenbroek, Michael, Hannes Grobe, Manfred Reinke, Uwe Schindler, Reiner Schlitzer, Rainer Sieger, und Gerold Wefer. 2002. „PANGAEA—an Information System for Environmental Sciences.“ *Computers & Geosciences* 28 (10): 1201–10. https://doi.org/10.1016/S0098-3004(02)00039-0.

Diepenbroek, Michael, Uwe Schindler, Robert Huber, Stéphane Pesant, Markus Stocker, Janine Felden, Melanie Buss, und Matthias Weinrebe. 2017. „Terminology Supported Archiving and Publication of Environmental Science Data in PANGAEA.“ *Journal of Biotechnology* 261: 177–86. https://doi.org/10.1016/j.jbiotec.2017.07.016.

 ISSN: 1860-7950

Friedrich, Tanja. 2020. „Looking for Data.“ Dissertation, Berlin: Humboldt Universität zu Berlin. https://doi.org/10.18452/22173.

Google Dataset Search. 2020. „Dataset Search: Metadata for Datasets.“ https://www.kaggle.com/datasets/1d97de37cb96c2c62182d22bf5924e0371933002bb7e5c46ba205b8a88de2e21.

Gregory, Kathleen, Helena Cousijn, Paul Groth, Andrea Scharnhorst, und Sally Wyatt. 2020. „Understanding Data Search as a Socio-Technical Practice.“ *Journal of Information Science* 46 (4): 459–75. https://doi.org/10.1177/0165551519837182.

Gregory, Kathleen, Paul Groth, Helena Cousijn, Andrea Scharnhorst, und Sally Wyatt. 2019. „Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines.“ *Journal of the Association for Information Science and Technology* 70 (5): 419–32. https://doi.org/10.1002/asi.24165.

Gregory, Kathleen, Paul Groth, Andrea Scharnhorst, und Sally Wyatt. 2020. „Lost or Found? Discovering Data Needed for Research.“ *Harvard Data Science Review* 2 (2). https://doi.org/10.1162/99608f92.e38165eb.

Ibáñez, Luis-Daniel, und Elena Simperl. 2022. „A Comparison of Dataset Search Behaviour of Internal Versus Search Engine Referred Sessions.“ In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, 158–68. New York: Association for Computing Machinery. https://doi.org/10.1145/3498366.3505821.

Kacprzak, Emilia, Laura Koesten, Luis-Daniel Ibáñez, Tom Blount, Jeni Tennison, und Elena Simperl. 2019. „Characterising Dataset Search—An Analysis of Search Logs and Data Requests.“ *Journal of Web Semantics* 55: 37–55. https://doi.org/10.1016/j.websem.2018.11.003.

Kacprzak, Emilia, Laura Koesten, Luis-Daniel Ibáñez, Elena Simperl, und Jeni Tennison. 2017. „A Query Log Analysis of Dataset Search.“ In *Web Engineering*, edited by Jordi Cabot, Roberto De Virgilio, und Riccardo Torlone, 429–36. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-319-60131-1_29.

Kern, Dagmar, und Brigitte Mathiak. 2015. „Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?“ In *Research and Advanced Technology for Digital Libraries*, edited by Sarantos Kapidakis, Cezary Mazurek, und Marcin Werla, 197–208. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-319-24592-8_15.

Khalsa, SiriJodha, Peter Cotroneo, und Mingfang Wu. 2018. „A Survey of Current Practices in Data Search Services.“ https://doi.org/10.17632/7j43z6n22z.1.

Koesten, Laura M., Emilia Kacprzak, Jenifer F. A. Tennison, und Elena Simperl. 2017. „The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour.“ In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1277–89. New York: Association for Computing Machinery. https://doi.org/10.1145/3025453.3025838.

Krämer, Thomas, Andrea Papenmeier, Zeljko Carevic, Dagmar Kern, und Brigitte Mathiak. 2021. „Data-Seeking Behaviour in the Social Sciences.“ *International Journal on Digital Libraries* 22 (2): 175–95. https://doi.org/10.1007/s00799-021-00303-0.

Kunze, Sven, und Sören Auer. 2013. „Dataset Retrieval.“ In, 1–8. https://doi.org/10.1109/ICSC.2013.12.

 ISSN: 1860-7950

Luk, Robert. 2022. „Why Is Information Retrieval a Scientific Discipline?“ *Foundations of Science* 27 (2): 427–53. https://doi.org/10.1007/s10699-020-09685-x.

Masson, Arnaud, Guido De Marchi, Bruno Merin, Maria Sarmiento, David Wenzel, und Beatriz Martinez. 2021. „Google Dataset Search and DOI for Data in the ESA Space Science Archives.“ *Advances in Space Research* 67 (8): 2504–16. https://doi.org/10.1016/j.asr.2021.01.035.

Meadow, Charles, Bert Boyce, Donald Kraft, und Carol Barry. 2007. *Text Information Retrieval Systems*. Cambridge, MA: Academic Press.

Noy, Natasha, und Omar Benjelloun. 2020. „An Analysis of Online Datasets Using Dataset Search.“ *Google AI Blog*. http://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html.

Roy, Dwaipayan, Zeljko Carevic, und Philipp Mayr. 2022. „Studying Retrievability of Publications und Datasets in an Integrated Retrieval System.“ In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 1–9. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3529372.3530931.

Sanderson, Mark, und Bruce Croft. 2012. „The History of Information Retrieval Research.“ *Proceedings of the IEEE* 100 (Special Centennial Issue): 1444–51. https://doi.org/10.1109/JPROC.2012.2189916.

Sun, Guangyuan, Tanja Friedrich, Kathleen Gregory, und Brigitte Mathiak. 2022. „Are We Building the Data Discovery Infrastructure Researchers Want? Comparing Perspectives of Support Specialists und Researchers.“ arXiv. https://doi.org/10.48550/arXiv.2209.14655.

Wang, Xu, Zhisheng Huang, und Frank van Harmelen. 2020. „Evaluating Similarity Measures for Dataset Search.“ In *Web Information Systems Engineering – WISE 2020*, edited by Zhisheng Huang, Wouter Beek, Hua Wang, Rui Zhou, und Yanchun Zhang, 38–51. Springer. https://doi.org/10.1007/978-3-030-62008-0_3.

Wu, Mingfang, Fotis Psomopoulos, Siri Jodha Khalsa, und Anita de Waard. 2019. „Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories.“ *Data Science Journal* 18 (1): 3. https://doi.org/10.5334/dsj-2019-003.

Dorothea Strecker ist Wissenschaftliche Mitarbeiterin im Projekt re3data COREF am Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin.

 ISSN: 1860-7950