COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

What are Researchers' Needs in Data Discovery? Analysis and Ranking of a Large-Scale Collection of Crowdsourced Use Cases

RESEARCH PAPER

]u[ubiquity press

BRIGITTE MATHIAK
NICK JUTY
ALESSIA BARDI
JULIEN COLOMB
PETER KRAKER

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data discovery is important to facilitate data re-use. In order to help frame the development and improvement of data discovery tools, we collected a list of requirements and users' wishes. This paper presents the analysis of these 101 use cases to examine data discovery requirements; these cases were collected between 2019 and 2020. We categorized the information across 12 'topics' and eight types of users. While the availability of metadata was an expected topic of importance, users were also keen on receiving more information on data citation and a better overview of their field. We conducted and analysed a survey among data infrastructure specialists in a first attempt at ranking the requirements. Between these data professionals, these rankings were very different, excepting the availability of metadata and data quality assessment.

CORRESPONDING AUTHOR:

Brigitte Mathiak

GESIS- Leibniz Institute for the Social Sciences, Germany

brigitte.mathiak@gesis.org

KEYWORDS:

Data Discovery; Research Data; Use Cases; Requirements

TO CITE THIS ARTICLE:

Mathiak, B, Juty N, Bardi, A, Colomb, J and Kraker, P. 2023. What are Researchers' Needs in Data Discovery? Analysis and Ranking of a Large-Scale Collection of Crowdsourced Use Cases. *Data Science Journal*, 22: 3, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2023-003

INTRODUCTION

Findability is at the core of the FAIR principles, yet researchers often report that data discovery, the finding of relevant data, is difficult (Chapman et al. 2020; Kern & Mathiak 2015). These difficulties are not only grounded in the question of metadata quality, but also involve a diverse array of supporting infrastructures and tooling. As of July 2020, re3data identified 2542 data depositories worldwide containing overlapping content. Consequently, data discovery cannot be accomplished by checking individual repositories, and more effort is needed to discover data, rather than scholarly literature, for example (Curty 2016). Even within a single data repository, finding high-quality research data with sufficient documentation for re-use can be challenging because research data, as opposed to research publications, is rarely peer-reviewed (Callaghan 2015). In a rapidly expanding ecosystem of research data, researchers are quickly overwhelmed by the flood of data.

All of this is detrimental to the re-use of research data; we know that only 15% of research data has been cited (Peters et al. 2016). While there are also other factors, the ease of discovery plays an important role in researchers' willingness to re-use data (Darby et al. 2012; Faniel, Majchrzak 2013). Discoverability is therefore one of the key challenges for FAIR data. In many ways, we cannot deliver on the promises of this movement if we do not increase the visibility of these research outputs.

RELATED WORK

There is still relatively little literature on the topic of scientific data discovery. This is defined as the whole process, starting with the researchers defining a data need to begin assessing the datasets they have found. Most works on data re-use (e.g. Tenopir 2011) focus on why and how people share data, and not why and how people re-use it. But as more research data is available, there has been a shift in the literature.

Chapman et al. (2019) offers comprehensive state-of-state analysis of dataset searches, which identifies open issues and particularly challenges how datasets should be queried and interlinked. While the aforementioned analysis is very comprehensive with respect to technical aspects, use cases and the user in general, political aspects of the research data landscape are only marginally considered.

Gregory et al. (2020b) reviewed 400 publications on data search and data discovery, focusing on user experience, especially with regards to discipline-specific differences and user needs. They understand data retrieval as 'a complex socio-technical process', given the importance of personal networks, literature, and collaborations in the process. This study suggests the significance of personal networks, literature and collaborations in data discovery, which is also supported by the findings in Friedrich (2020), Marée (2016), Yoon (2014), and Krämer et al. (2021).

For this work, we offer a third perspective on the process: how these user needs are translated into use cases, and how these use cases are perceived and ranked by those people who run data discovery services themselves. This use case collection is not the first of its kind; the RDA Interest Group on Data Discovery Paradigms collected 66 use cases (de Waard et al. 2017), which are also included in our collection. From this, they derived recommendations for data repositories (Wu et al. 2019). We have also included use cases inspired by user studies (e.g. Krämer et al. 2021) to include the user perspective, taking into account feedback from experts in the field as well as scientists.

PROBLEM STATEMENT

Infrastructures play a pivotal role in enabling data discovery. In an effort to determine whether existing infrastructures are sufficient enough to support this crucial function, we have gathered a number of use cases from a wide spectrum of disciplines and sources to identify omissions and inadequacies that could impact this functionality. In this paper, we offer an updated and expanded view of use cases, as well as how these use cases are viewed by research data professionals. For this, we expanded the number of use cases to 101. Since it was impractical for research data professionals to individually rank this number of use cases, we first grouped the use cases into 12 clusters, such as 'data citation', 'machine discoverability', and 'linking'. We then asked 25 data professionals to identify the two most important and the two least important clusters from their point of view, based on their job role and regular work requirements.

METHODOLOGY

DATA COLLECTION

The use cases themselves follow this format: As a [role] I want to [goal] so that [benefit]. Most of the use cases were identified by Implementation Network (IN) members—including researchers, librarians, and infrastructure providers—or documented during focused events by participating contributors. One such event was the workshop 'Data Discovery Across Disciplines' at Open Science Fair 2019, where members of the audience were asked to contribute. Another important source was the use case collection of the RDA IG Data Discovery. We have provided a URL for the sources whenever possible. Altogether, 101 use cases[1] were collected. The use cases, their ranking, the survey data itself, and the analysis is available at https://doi.org/10.5281/zenodo.5006524 (Mathiak et al. 2021). Tables 1 and 2 below provide more information as to the structure and content of the use case data sheet.

Data Description

Table 1 Table description. Description of each column in the use case table, documenting the information intended to be captured from use cases.

COLUMN	DESCRIPTION
A	ID of the use cases; these are referenced below and immutable to any sorting
B	Use Case: The use case in the form "As a [role] I want to [goal] so that [benefit]"
C	Actor: The role perspective of the use case, copied out for your convenience
D	Cluster: The name of the cluster (see *Method* and *Cluster* sections below)
E	Contributed by: Gives you the name of the IN member who entered this use case.
F	Source: The event or source this use case was found (see table below)
G	Closely_related_to: ID of use cases that were related to this one, nearly duplicates

Table 2 Documentation of the sources used.

SHORTHAND FOR SOURCE	DESCRIPTION (INCL. WEBLINK, IF AVAILABLE)
Implementation Network	These use cases were found by the members of the Implementation Network. https://www.go-fair.org/implementation-networks/overview/discovery/
Observation Study	These use cases were taken from Krämer et al. (2021)
Workshop	'Data Discovery Across Disciplines' at the Open Science Fair 2019. During the session, participants were asked for their use cases. https://www.go-fair.org/2019/11/04/workshop-report-data-discovery-across-disciplines-at-open-science-fair-2019/
RDA IG Data Discovery	The RDA interest group has collected these use cases through a survey. https://doi.org/10.5281/zenodo.1050976
EUDAT-Prace Summer School 2019	These use cases were collected during the summer school. https://eudat.eu/eudat-prace-summer-school-2019
DKRZ Researcher	These use cases were collected from individual researchers working at the DKRZ (https://www.dkrz.de/), the German Climate Computing Centre.

CLUSTERING

After collecting all the use cases, we proceeded to annotate them with regards to the types of entities they referred to, e.g. datasets, papers, persons, etc. Using these annotations as a guideline, two of the authors clustered the use cases thematically, aiming to provide an overview, identify common themes, and facilitate identification of direct duplicates or equivalent use cases.

[1] https://doi.org/10.5281/zenodo.5006524.

We found six use cases which were almost identical, differing only in the actor involved, for example student versus researcher (38 and 89, 57 and 96, 87 and 30, 39 and 90, 88 and 35, 40 and 91, see the Closely_related_to column in the source data). Indeed, most student cases were also valid for researchers, but we decided to keep the two categories apart to retain the original use case wording. We also excluded five use cases (110, 23, 27, 25 and 77) from consideration, because we felt they were not data discovery use cases. They are, however, included in the use case documentation as they may offer additional motivations for data discovery in the future.

Our motivation for clustering was to put together use cases that would require similar functionality from the data infrastructure, and may therefore provide inspiration with regards to how the infrastructure may evolve. We tried to envision how the use cases could be operationalized. In rare cases, where there was ambiguity or overlap in the use case assignment to a cluster, we assigned that use case to two clusters.

RANKING ANALYSIS OF THE CLUSTERS

After collating and categorising the use cases that had been submitted, as described above, we conducted a survey to ascertain their relative importance; to elicit a community consensus, we used a survey tool during the GOFAIR meeting on February 3rd, 2021. Participants were asked to rank the two most and the two least 'relevant' use case clusters, in addition to providing routine demographic and role information.

In conducting the participant information survey (25 responses out of a total of about 50 attendees, data also available at 10.5281/zenodo.5006524), we found that most participants hailed from mainland Europe (17), with 3 additional participants from the UK. Of those mainland participants, most were located in Germany (8), Scandinavian countries (4), and the Netherlands (3). There were 3 attendees from North America (2 US, 1 Canada), and 1 South American attendee (Brazil). The participant roles of these attendees were very skewed towards infrastructure providers (17), with 8 respondents providing data themselves. Only 7 people said they actually searched for data, though that was never the only selection for that question. As such, we can conclude that most of the respondents have a data infrastructure background, and while they may not necessarily be in the position to make policy decisions for data infrastructures, they nonetheless have a good understanding of the priorities and policies of the infrastructure institutions they work for.

Most respondents were involved in life science (8), general science (7), or natural science (5). The social science (3) and computer science domains (1) were not well represented. Interestingly, when asked which tools were used besides a typical 'Google' search, by far the most popular answer was the use of domain or community specific tools (10 of 21 responses). The next most popular response written in the form of free-text answers were publications (3), DataCite (3), and colleagues (2). All other answers were individual responses. While the respondents were largely drawn from a biological/life sciences background, this did highlight the common use of domain-specific tooling.

For the ranking analysis, we are using descriptive statistics to identify clusters that were ranked by many participants as relevant or irrelevant. Additionally, we gauge general interest in a cluster by how often it was picked in either category. Controversial clusters, such as those that were picked as important by some and irrelevant by other participants can point towards differences in domains or different paradigms within the community. While all participants chose a most relevant and second-most relevant cluster, there were 5 incomplete or 'I don't know' answers for least relevant and also another 8 for not relevant.

RESULTS

One of the largest clusters we found (14 ideas, counting duplicates only once) surprisingly was on **data citations**: finding papers that cite the dataset, finding datasets that were frequently cited, finding not only papers, but also other media (like slides, websites and blogs) that use and/or cite a certain dataset. This is surprising not only because of the inventiveness demonstrated in the use of these citations (14 ideas in total), but also because there are relatively few services that support these use cases and a scarcity of data in all domains, despite citation being a common and well-established scientific practice.

Another large cluster (14 ideas) of use cases, the **overview**, focused on users' needs, particularly in terms of finding data by topic or research question, as well as helping them with re-use. However, this cluster does not have many supporting infrastructure providers. Specific topics, such as micro-plastic and satellites, were used as case examples, which could be extrapolated to other domains and highly granular topics.

There were some use cases that would ask for additional **metadata** to enable **discovery** (12 ideas, counting duplicates only once). This includes many of the common metadata fields (such as licenses, number of data points, and format), but also makes a case for rarely used or specialized metadata fields such as experimental design, publication date, and the precise instrument that was used to obtain/generate the data. Ideally, metadata should also be multilingual and use identifiers to facilitate cross-referencing.

Documentation, in the form of additional documents, is the focus of many use cases (9 ideas). It is invaluable for re-use and correct interpretation, and it should be kept closely linked to the dataset itself. Documentation does have some overlap with the data citation cluster, as documentation often comes in the form of scientific publications that have used the dataset before.

Many use cases addressed the **discoverability** of the datasets (8 ideas). Here, where given a choice, users wanted to put their data into the repository with the highest discoverability, and get feedback on the impact of their dataset—for instance its re-use and by whom, where possible. The desire to add annotations to improve the discoverability of specific datasets was strongly expressed. Developers wanted to maximize discoverability and one actor wanted to make sure that everyone, including researchers from poorer nations, would be able to find research data.

Convenience when using discovery systems was another cluster of use cases (8 ideas). In this cluster, we collected use cases that would make the life of users easier, such as the visualization of molecules, thumbnails, RSS-type notification, and autocomplete functionality. There are use cases for both import and export functions of dataset entries, reminding us that not all users are consuming data — some are trying to input data as well.

Other topics (6 use cases each) were: **Linking** (1) between datasets, both identical and related; **Quality** (2) focused on metadata expansion to incorporate organizations or data providers, and assess trustworthiness, and status with respect to peer-review. **Cross-domain** (3), for enabling discovery between domains, such as through cross-domain meta search engines, links between topics, and crosswalks between different vocabularies.

Searching **within the dataset** (5 ideas) is important for large datasets, e.g. the ones provided by the government, in biodiversity research, or for climate observation.

Two smaller clusters that we nevertheless found quite interesting were **Person** (4 ideas) and **Machine discoverability** (2 entries). Use cases in the cluster 'Person' referred to persons as entities, and aim e.g. finding persons that would generate datasets of a specific kind, or as a combination with data citation, where people have published on a specific dataset. Machine discoverability was mentioned in two use cases, with the general goal to run algorithms on the dataset metadata.

The main actors in our use case collection are researchers and students, who have very similar, if not identical needs. There are two clusters that are different in this regard: **Discoverability** of research data is something that seems to concern data producers more than the researchers. This makes sense, given that data producers have a vested interest in being seen, cited, and re-used. The other cluster is **search within the dataset**, which is skewed towards domain-specific researchers. It should be noted that these are very different domains from very different fields.

For the ranking of documented use cases, we pooled the two high and low ranking answers into one category each, and plotted the number of times each cluster was was placed into either the high or low categories (Figure 1). Interestingly, nearly all use case categories were designated at least once in the low and high ranking groups respectively (apart from the overview category). There was a relatively high consensus for placing Discoverability and

Metadata for quality assessment high in the ranking list, while searching for data in a dataset or looking for specific researchers were both given a low ranking with the majority of the audience. Unfortunately, we did not link the two surveys and cannot decipher if different target groups might have had different rankings.

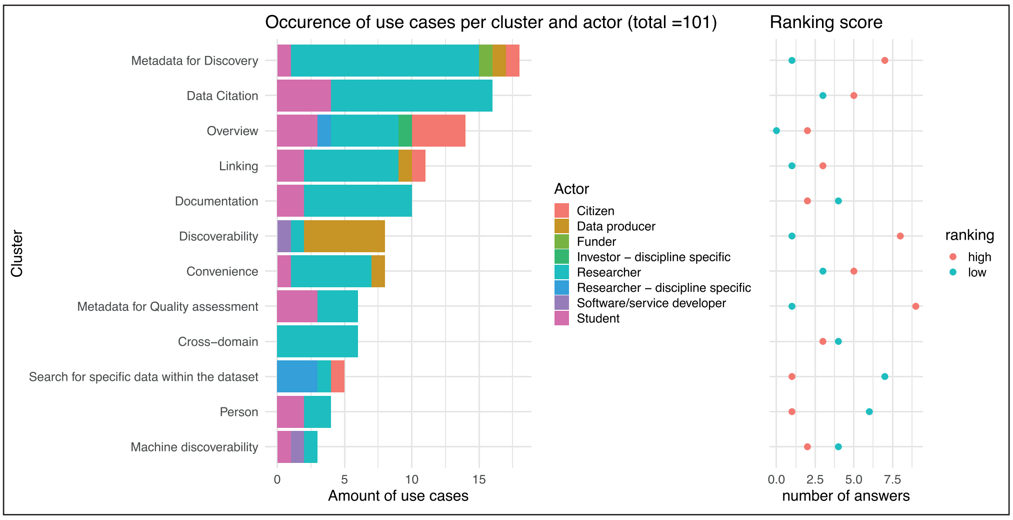

Figure 1 Summary of use case distribution (left) and ranking (right). Distribution is plotted depending on use case categories (cluster) and actor. Six use cases were reported for both researchers and students, and some use cases proposed for students may be relevant for researchers. The ranking of each category was assessed in a survey with 25 answers.

DISCUSSION

We present here 101 use cases, categorised as 12 clusters and 8 'actors'. This is currently the most complete collection of such use cases. However, it should be noted that the main contributors to this list of use cases were experts in the field, representing people who are familiar with information infrastructures rather than the main users of such systems, specifically scientists themselves. This bias is partially addressed through the inclusion of use cases from user studies and use cases added by scientists, but those constitute a minority overall.

This bias manifests again more strongly in the survey where questions were directed solely at data professionals. An additional bias also exists in the fact that participant origin was heavily skewed towards the European Union countries, as we have found. We anticipate that this likely reflects the representative infrastructural/ecosystem state of most Western countries, but it may not be indicative of the same for poor nations. In addition, we acknowledge that this information is drawn from a limited pool of participants and from limited domains, being largely centred on life science.

Comparing the results of the ranking to the needs of scientists, as identified in user studies, we found that personal networks, literature, and collaborations are seen as quite important (Friedrich 2020; Marée 2016; Yoon 2014; Krämer et al. 2021), while in the user studies, the corresponding clusters were not highly prioritised. This is of course, due to the different perspectives of these two groups of people. Users are unlikely to complain about missing or bad metadata, as they are unlikely to ever encounter such things due to ranking. Data professionals on the other hand, are quite aware of the problem and how it makes certain datasets de facto invisible, thereby removing them from the cycle of data re-use.

An alternative explanation could be that data professionals are often coming from a library background (Cox et al. 2019) and have a tendency to treat research data the same way they would literature, using the same infrastructure and methodology. Differences between those two entity types tend to get swept away by the practicalities of having to provide productive search environments, as well as the lack of viable alternatives. While the user studies point towards differences, they do not offer a constructive way to do better, nor provide empirical evidence that this hypothetical 'better way' would *actually* perform better.

CONCLUSION

Data discovery has been described by users as difficult (Chapman et al. 2020) in comparison to searching for literature. In determining the best way forward — and how to best allocate available resources to address this discrepancy — it is tempting to replicate the existing infrastructure for literature, but this may not be the best way forward.

Data discovery is limited by the existing infrastructure ecosystem; improving data discovery would therefore require both the augmentation of existing systems, as well as a more well planned and implemented infrastructure in the future. With this in mind, we anticipate that the use cases we have collected will help drive the augmentation of existing infrastructure, as well as assisting with the appropriate planning and development of future infrastructure. We present here a multi-method study and analysis of more than 101 use cases for data discovery requirements, identifying, collating and categorising themes or topics from a variety of actors.

In comparison to the current literature on user studies, we identified gaps between what is typically identified as important and how the clusters were actually ranked by data professionals. Our attempt at ranking these categories showed an absence of overall consensus, even though most participants in this work represented a unique actor, namely 'infrastructure providers'. This suggests that a more precise analysis, probably limited to specific target groups (and perhaps even sub-groups) might be needed to rank the use cases. On the other hand, it may well be that the whole spectrum is important and that one would need to satisfy most requests simultaneously to create tools/infrastructure universally appropriate for discovery.

The outcomes of this work do however, point to significant gaps in existing search infrastructure, where interconnections between 'person', 'dataset', and other entities are lacking, especially at the level of 'citation' and interconnections between these entities.

We anticipate that the gaps identified in this work are highly generalizable across existing infrastructures, as well as across domains, and can be used to devise plans to improve those infrastructures and guide the development of a more integrated infrastructure ecosystem, particularly for emerging and nascent systems.

We know of many existing activities that seek to close these gaps, e.g. Scholix for the connection between datasets and literature, and OpenAIRE, which connects datasets and authors with citations and recognize the fast pace at which the ecosystem is currently evolving. To future readers, who may be puzzled by our results, we would like to point out that this work represents a snapshot of the situation as it presented itself around 2019 to 2022, and may hopefully be made obsolete in the near future.

ACKNOWLEDGEMENTS

We would like to offer our thanks to the authors of the dataset (10.5281/zenodo.5006525) (Mathiak et al. 2021) and other members of the GO FAIR Discovery IN (https://www.go-fair.org/implementation-networks/overview/discovery/), for collecting and structuring the use cases, as well as providing feedback. We would like to acknowledge the participants of the workshop 'Data Discovery Across Disciplines' at Open Science Fair 2019, who provided us with further use cases, the RDA Discovery Paradigm WG, and other people who provided us with valuable feedback as well as additional use cases.

FUNDING INFORMATION

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of NFDI- 442494171 and the European Commission as part of the H2020 projects TRIPLE (grant agreement no. 863420), OpenAIRE-Advance (grant agreement no. 777541).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Brigitte Mathiak orcid.org/0000-0003-1793-9615
GESIS- Leibniz Institute for the Social Sciences, Germany

Nick Juty orcid.org/0000-0002-2036-8350
The University of Manchester (ELIXIR-UK), United Kingdom

Alessia Bardi orcid.org/0000-0002-1112-1292
CNR-ISTI (CNR-ISTI), Italy

Julien Colomb orcid.org/0000-0002-3127-5520
Humboldt-Universität zu Berlin (HU Berlin), DE

Peter Kraker orcid.org/0000-0002-5238-4195
Open Knowledge Maps, Austria

REFERENCES

Callaghan, S. 2015. Data without peer: Examples of data peer review in the Earth sciences. *D-Lib Magazine,* 21(1): 9. DOI: https://doi.org/10.1045/january2015-callaghan

Chapman, A, Simperl, E, Koesten, L, Konstantinidis, G, Ibáñez, L-D, Kacprzak, E and **Groth, P.** 2020. Dataset search: a survey. *The VLDB Journal,* 29(1): 251–272. DOI: https://doi.org/10.1007/s00778-019-00564-x

Cox, AM, Kennan, MA, Lyon, L, Pinfield, S and **Sbaffi, L.** 2019. Maturing research data services and the transformation of academic libraries. *Journal of Documentation*, 75(6): 1432–1462. DOI: https://doi.org/10.1108/JD-12-2018-0211

Curty, RG. 2016. Factors influencing research data reuse in the social sciences: An exploratory study. *International Journal Digital Curation,* 11(1): 96–117. DOI: https://doi.org/10.2218/ijdc.v11i1.401

Darby, RS, Lambert, BM, Wilson, M, Gitmans, K, Dallmeier-Tiessen, S, Mele, S and **Suhonen, J.** 2012. Enabling scientific data sharing and re-use. *2012 IEEE 8th International Conference on E-Science*. IEEE. DOI: https://doi.org/10.1109/eScience.2012.6404476

de Waard, A, Khalsa, SJ, Psomopoulos, F and **Wu, M.** 2017. RDA IG Data Discovery Paradigms IG: Use Cases data [Data set]. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.1050976

Faniel, IM, Barrera-Gomez, J, Kriesberg, A and **Yakel, E.** 2013. A comparative study of data reuse among quantitative social scientists and archaeologists. In: *iConference*. DOI: https://doi.org/10.9776/13391

Friedrich, T. 2020. Looking for data. DOI: https://doi.org/10.18452/22173

Gregory, KM, Cousijn, H, Groth, P, Scharnhorst, A and **Wyatt, S.** 2020b. Understanding data search as a socio-technical practice. *Journal of Information Science*, 46(4): 459–475. DOI: https://doi.org/10.1177/0165551519837182

Kern, D and **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*. DOI: https://doi.org/10.1007/978-3-319-24592-8_15

Krämer, T, Papenmeier, A, Carevic, Z, Kern, D and **Mathiak, B.** 2021. Data-Seeking Behaviour in the Social Sciences. *International Journal on Digital Libraries,* 1–21. DOI: https://doi.org/10.1007/s00799-021-00303-0

Mathiak, B, Juty, N, Heger, T, Di Donato, F, Jeschke, J, Widmann, H, Kraker, P, et al. 2021. Stocktaking GO FAIR Discovery IN - Use cases, infrastructure (Version 0.9) [Data set]. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5006525

Peters, I, Kraker, P, Lex, E, Gumpenberger, C and **Gorraiz, J.** 2016. Research data explored: an extended analysis of citations and altmetrics. *Scientometrics,* 107: 723. DOI: https://doi.org/10.1007/s11192-016-1887-4

re3data.org. Registry of Research Data Repositories. DOI: https://doi.org/10.17616/R3D last accessed: 2020-07-16.

Tenopir, C, Allard, S, Douglass, K, Aydinoglu, AU, Wu, L, Read E, et al. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6(6): e21101. DOI: https://doi.org/10.1371/journal.pone.0021101

Wu, M, Psomopoulos, F, Khalsa, SJ and **de Waard, A.** 2019. Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. *Data Science Journal*, 18(1): 3. DOI: https://doi.org/10.5334/dsj-2019-003

Yoon, A. 2014. Making a square fit into a circle: Re-searchers' experiences reusing qualitative data. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–4. DOI: https://doi.org/10.1002/meet.2014.14505101140

TO CITE THIS ARTICLE:
Mathiak, B, Juty N, Bardi, A, Colomb, J and Kraker, P. 2023. What are Researchers' Needs in Data Discovery? Analysis and Ranking of a Large-Scale Collection of Crowdsourced Use Cases. *Data Science Journal*, 22: 3, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2023-003

Submitted: 24 September 2021
Accepted: 19 January 2023
Published: 09 February 2023

COPYRIGHT:
© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.