













Connecting Researchers to Data Repositories in the Earth, Space, and Environmental Sciences

Michael Witt¹ , Shelley Stall² , Ruth Duerr³ ,
Raymond Plante⁴ , Martin Fenner⁵ , Robin Dasler⁵ ,
Patricia Cruse⁵ , Sophie Hou⁶ , Robert Ulrich⁷ ,
and Danie Kinkade⁸ 

¹ Purdue University, West Lafayette, IN, USA
mwitt@purdue.edu

² American Geophysical Union, Washington DC, USA
sstall@agu.org

³ Ronin Institute, Montclair, NJ, USA
ruth.duerr@ronininstitute.org

⁴ National Institute of Standards and Technology, Gaithersburg, MD, USA
raymond.plante@nist.gov

⁵ DataCite, Hannover, Germany
{martin.fenner, robin.dasler,
patricia.cruse}@datacite.org

⁶ National Center for Atmospheric Research,
University Corporation for Atmospheric Research, Boulder, CO, USA
hou@ucar.edu

⁷ Karlsruher Institut für Technologie, Karlsruhe, Germany
robert.ulrich@kit.edu

⁸ Woods Hole Oceanographic Institution, Woods Hole, MA, USA
dkinkade@whoi.edu

Abstract. The Repository Finder tool was developed to help researchers in the domain of Earth, space, and environmental sciences to identify appropriate repositories where they can deposit their research data and to promote practices that implement the FAIR Principles, encouraging progress toward sharing data that are findable, accessible, interoperable, and reusable. Requirements for the design of the tool were gathered through a series of workshops and working groups as a part of the Enabling FAIR Data initiative led by the American Geophysical Union that included the development of a decision tree that researchers may follow in selecting a data repository, interviews with domain repository managers, and usability testing. The tool is hosted on the web by DataCite and enables a researcher to query all data repositories by keyword or to view a list of domain repositories that accept data for deposit, support open access, and provide persistent identifiers. Metadata records from the re3data.org registry of research data repositories and the returned results highlight repositories that have achieved trustworthy digital repository certification through a formal procedure such as the CoreTrust Seal.

Keywords: Research data management · FAIR principles · Geosciences · Repositories · Data facilities · Recommender systems

1 Background

The Enabling FAIR Data initiative¹ was organized by the American Geophysical Union (AGU) with a goal of promoting the adoption of the FAIR Principles: sharing research datasets in the domain of Earth, space, and environmental sciences and making them findable, accessible, interoperable, and reusable [1]. It accomplished its work through a series of two stakeholder meetings, two workshops, and six working groups (called targeted adoption groups, or TAGs) that engaged over three hundred researchers, publishers, funders, professional societies, repository managers, librarians, and other data professionals over the course of twelve months between November 2017 and October 2018.

There were many notable outcomes from the initiative and the work of the TAGs. Building on the past efforts of the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS), the “Enabling FAIR Data Commitment Statement in the Earth, Space, and Environmental Sciences”² was drafted and outlined responsibilities for repositories; publishers; societies, communities, and institutions; funding agencies and organizations; and individual researchers to sign and commit to strive for FAIR data in their domain. A set of guidelines were developed for journals and publishers³ to instruct their authors to deposit data in FAIR-aligned repositories (e.g., instead of including data as supplementary files with their articles); to cite and link to data in their articles; to include a data availability statement; and to provide open access to data that support published findings except in cases of ethical or legal constraints. The Data Management Training Clearinghouse⁴ began collecting information about educational materials related to research data management and the data lifecycle through a novel approach of crowdsourcing events; this resource will continue to be developed and maintained through the Earth Science Information Partners (ESIP) with a subsequent grant from the Institute of Museum and Library Services. Other efforts included the definition of workflows around persistent identifiers such as Digital Object Identifiers (DOIs) between data repositories and publishers [2]; discussion and outreach related to creating a change in culture by giving proper credit to those who enable FAIR data; and the establishment of a program with funds to support a cohort of repositories in the Earth, space, and environmental sciences to pursue CoreTrustSeal certification⁵ and become formally recognized as trustworthy digital repositories.

In this context, the Repository Guidance TAG was motivated by the use case of a researcher who is producing data and trying to identify appropriate domain repositories

¹ <http://www.copdess.org/enabling-fair-data-project>.

² <http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences>.

³ <http://www.copdess.org/enabling-fair-data-project/author-guidelines>.

⁴ <http://dmclearinghouse.esipfed.org>.

⁵ <https://www.coretrustseal.org>.

that will accept their data for deposit. A secondary motivation was to promote practices that implement or otherwise support the FAIR Principles, in particular, among data repositories. Through a series of online discussions and work in person, the group diagrammed the decisions that a researcher commonly faces when selecting a repository to deposit their data. An interview protocol was also developed, and interviews were conducted with eleven repository managers to better understand the current state of domain repositories in implementing or planning to implement the FAIR Principles. Requirements were derived from these two activities to design an online, web-based tool that would be simple and easy for researchers to identify and select an appropriate repository. The tool, Repository Finder⁶, queries the re3data registry of research data repositories⁷ using Elasticsearch. Wireframing and software development took place in the final six months of the initiative in collaboration with the AGU and DataCite. Repository Finder was inspired, in part, by the Digital Research Infrastructure for the Arts and Humanities (DARIAH) Data Deposit Recommendation Service that provides a similar tool for European arts and humanities researchers that enables users to select their research discipline and country from drop-down lists and presents a list of repositories in their domain and locale that may accept their data [3]. A prototype of the Repository Finder was introduced, and three usability studies resulted in feedback from users that was incorporated into the tool. Although challenges and constraints were met throughout the process, the tool has been well-received by the Earth, space, and environmental sciences research community, and opportunities exist for future work in expanding the tool to cover other domains and extending its functionality to meet additional needs that were expressed by users during usability testing and early adopters of the tool in production.

2 Design and Development

2.1 Decision Tree

The starting place for the group was to imagine a researcher who is producing data and to diagram at a high level what decisions that researcher would make as they select a repository to deposit their data. Researchers could be writing proposals, for example, to funding agencies that require data management plans in which repositories must be specified; or, they could be further along in performing their research at a point when they are submitting papers for publication with journals that require supporting data be shared and archived in a repository (Fig. 1).

In all cases, researchers will be compelled to follow any mandates or recommendations of specific repositories that are made by their funder, publisher, journal, or institution. If this is not the case, researchers should consider what domain-specific repositories are commonly used in their area of research or for the type of data they are producing. They should contact the repository or otherwise confirm whether they can deposit their data and what the parameters for using the service are, e.g., what metadata

⁶ <http://repositoryfinder.datacite.org>.

⁷ <http://re3data.org>.

When selecting a repository to deposit their data, researchers in the Earth, space and environmental sciences should:

1. Follow the mandate or recommendation of their funder / publisher / journal / institution / etc.
2. Use a domain repository that adheres to norms of their research community*
3. Use a certified repository**
4. Use the repository with the highest level of curation*
5. Use an institutional repository*
6. Use a general repository

* that will take their data based on the repository's terms of use or direct interaction between the researcher and the repository staff

** leveraging this opportunity to encourage/facilitate repositories who are not certified to pursue certification, e.g., CoreTrustSeal

Fig. 1. Ordered list of principles distilled from the decision tree.

and formats do they require, do they charge a fee, file size limits, etc. A repository that is certified is preferable to one that is not because it has demonstrated through a formal process such as the CoreTrustSeal or ISO 16363 [4] that it has an adequate and sustainable organizational infrastructure, digital object management, and technology to provide the services that it advertises. Furthermore, repositories that provide higher levels of curation are preferable to repositories that provide lower levels of curation. In the context of CoreTrustSeal, four levels of curation are defined [5]. The lowest level of curation is to simply accept and distribute content as-is deposited. Basic curation may involve a simple check of the data before it is accepted and the addition of some basic metadata or documentation. Enhanced curation may additionally include format conversion and more detailed metadata or documentation. The highest level, or data-level, curation includes enhanced curation with editing and quality assurance of the data or other, additional services. If a domain repository is not available or is unable to accept the researcher's data, they may use a data repository that is offered by their institution or a general-purpose data repository such as Dryad⁸, figshare⁹, Harvard Dataverse¹⁰, Mendeley Data¹¹, Open Science Framework¹², and Zenodo¹³. Likewise for institutional or general-purpose data repositories, certification and higher levels of curation

⁸ <http://datadryad.org>.

⁹ <https://figshare.com>.

¹⁰ <https://dataverse.harvard.edu>.

¹¹ <https://data.mendeley.com>.

¹² <https://osf.io>.

¹³ <https://zenodo.org>.

are preferred. In many research organizations, researchers can consult librarians for assistance in navigating these options and for help in selecting an appropriate repository for their data [6], and further guidance is offered by resources such as the Digital Curation Centre's Checklist for Evaluating Data Repositories [7].

2.2 Interviews with Data Facilities

Initiated by the Lorentz Workshop¹⁴ in 2014, members of the broader research community collaborated to develop and publish the FAIR Principles in 2016 [1], and work has continued through the GO FAIR initiative¹⁵ towards defining metrics for evaluating and measuring implementation of the Principles or the “FAIRness” of data and metadata. To get a better, practical sense of FAIR adoption among data repositories in the Earth, space, and environmental sciences, the TAG designed an interview guide [8] and conducted one-hour interviews with domain repository managers who were engaged in the Enabling FAIR Data initiative. These included the Ag Data Commons (United States Department of Agriculture), Alabama Geological Survey, Biological and Chemical Oceanography Data Management Office (Woods Hole Oceanographic Institution), Dash (California Digital Library), Deep Carbon Observatory, Interdisciplinary Earth Data Alliance (Columbia University), Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, PANGAEA, Socioeconomic Data and Applications Center (NASA), Research Data Archive (National Center for Atmospheric Research), and VTechData (Virginia Polytechnic Institute and State University).

The questions were developed from discussions within the TAG (including those who would be interviewed) and reflect what they considered FAIR implementation to resemble in current repository practice that are salient within the domain. For each practice, repositories were asked if they had implemented the practice, had plans to implement it, or did not have plans for implementation. Common practices among repositories in the domain included providing a search/browse user interface; provisioning landing pages for datasets; minting Digital Object Identifiers (DOIs) for datasets and/or data collections; providing human-readable and machine-actionable metadata describing datasets; linking datasets to related, published literature; supporting interfaces for metadata export and harvesting; identifying authors using ORCID identifiers; describing datasets with temporal, geospatial, and other domain-specific metadata; suggesting citations to encourage users to cite data; providing support services around data (e.g., help desk); enabling direct machine access to data (e.g., ftp, THREDDS, OPeNDAP, SPARQL, OGC); ascribing open access licenses for data reuse; registering their repositories with re3data; and supporting functionality to include data housed by the repository in peer review workflows. All repositories recognized the importance of certification: approximately half were already certified (primarily through the World Data System¹⁶ or CoreTrustSeal) or were actively

¹⁴ <http://www.lorentzcenter.nl/lc/web/program.php3?jaar=2015>.

¹⁵ <https://www.go-fair.org>.

¹⁶ <https://www.icsu-wds.org>.

pursuing certification, with the other half either planning to or interested in pursuing certification in the near future. Only some repositories embed machine-actionable metadata in landing pages (e.g., JSON-LD, HTML meta tags); linked or otherwise referenced datasets in their repositories to related data elsewhere; captured provenance of data in their custody; or furnished citations to related literature in their DOI metadata, e.g., such that could be used by Scholix¹⁷ or other services that relate data and literature, quantify impact measures of data, etc. Interestingly, while most repositories responded that they do not provide machine-actionable citations for their data, these were effectively provided by the DOI for many applications such as citation management software clients.

2.3 re3data Schema Mapping and Tool Design

The primary purpose of the decision tree and interview exercises were to inform the design and development of the Repository Finder: to make it easy for a researcher to identify an appropriate domain repository to deposit their data and, in the process, to tacitly promote the FAIR Principles both in terms of awareness for the researcher and to begin to recognize emerging “FAIR” practices by data repositories (Fig. 2).

311-01 Astrophysics and Astronomy	311 Astrophysics and Astronomy	32 Physics	34 Geosciences (including Geography)	3 Natural Sciences
313-01 Atmospheric Science	313 Atmospheric Science, Oceanography and Climate Research			
313-02 Oceanography	314 Geology and Palaeontology			
314-01 Geology and Palaeontology	315 Geophysics and Geodesy			
315-01 Geophysics	316 Geochemistry, Mineralogy and Crystallography			
315-02 Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartography	317 Geography			
316-01 Geochemistry, Mineralogy and Crystallography	318 Water Research			
317-01 Physical Geography				
317-02 Human Geography				
318-01 Hydrogeology, Hydrology, Limnology, Urban Water Management, Water Chemistry, Integrated Water Resources Management				

Fig. 2. Relevant domain Deutsche Forschungsgemeinschaft (DFG) subject classifications.

The re3data registry currently manages metadata records describing 2,200 data repositories across all domains of research and around the world. Each repository is cataloged using forty-one descriptive attributes that are explained and can be validated in XML¹⁸ using version 3.0 of the Metadata Schema for the Description of Research Data Repositories [9]. A subset of records pertaining to the Earth, space, and environmental sciences was established by limiting to relevant subjectID attributes based on the Classification of Subject Area, Review Board, Research Area and Scientific Discipline (2016–2019) from the Deutsche Forschungsgemeinschaft (DFG)¹⁹, which is used by re3data. SubjectID is a required attribute in the schema. In addition, results are implicitly limited to only repositories that accept data for deposit (dataUploadType is “open” or “restricted”) and those that are domain repositories (type is “disciplinary”).

¹⁷ <http://www.scholix.org>.

¹⁸ <https://www.re3data.org/schema>.

¹⁹ http://dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/fk-wahl2015/2015_fachsystematik_2016_2019_en.pdf.

Guided by the interview and user test results, repositories that provide open access to their data (dataAccessType is “open”) and persistent identifiers (pidSystem is true) were included in the criteria for inclusion to recognize practices that are working towards FAIR that are currently and widely adopted (Fig. 3).

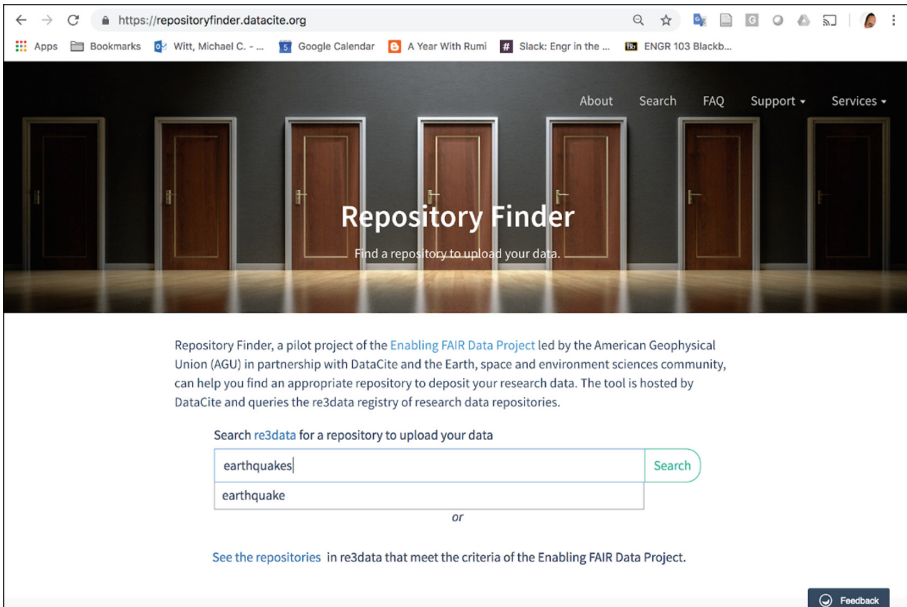


Fig. 3. Web-based user interface of the repository finder.

Initial design documents and discussions were incorporated into a wireframe using Balsamiq²⁰ that was iterated through biweekly online meetings of the project team. The software was developed as two separate applications: an API integrated with the re3-data Elasticsearch index using the Ruby on Rails²¹ framework, and a frontend for this API using the EmberJS²² framework. The work was done over the course of four monthly development sprints, and the tool is hosted by DataCite with its source code openly accessible on github^{23,24}. The application queries re3data using Elasticsearch match phrase prefix queries²⁵ on the repositoryName, description, and keyword fields. Repository Finder begins with a short explanation of the tool and then presents two

²⁰ <https://balsamiq.com>.

²¹ <https://rubyonrails.org>.

²² <https://www.emberjs.com>.

²³ <https://github.com/datacite/dackel>.

²⁴ <https://github.com/datacite/schnauzer>.

²⁵ <https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-match-query-phrase-prefix.html>.

options: the user can (1) initiate a search by entering keywords that are auto-completed and receive results ranked by relevance of all repositories that accept deposit of data, provide open access, and use persistent identifiers. Alternatively, the user can (2) click a link to see repositories that meet the criteria of the Enabling FAIR Data community. These results are ordered alphabetically and include only repositories in the domain of Earth, space, and environmental sciences that meet the above criteria, highlighting repositories that have achieved certification with a “seal” icon. After results are displayed, the user has the ability to narrow the results by keyword; subsequent results are ranked by relevance. Each individual result displays the name of the repository; its description, subjects, and keywords; and a switch to display more details about the repository, including links to the repository, contact information, and its full registry entry in the native re3data interface. At the end of the result list, institutional and general-purpose repositories are suggested to be used for cases where a domain repository is not available or will not accept the researcher’s data (Fig. 4).

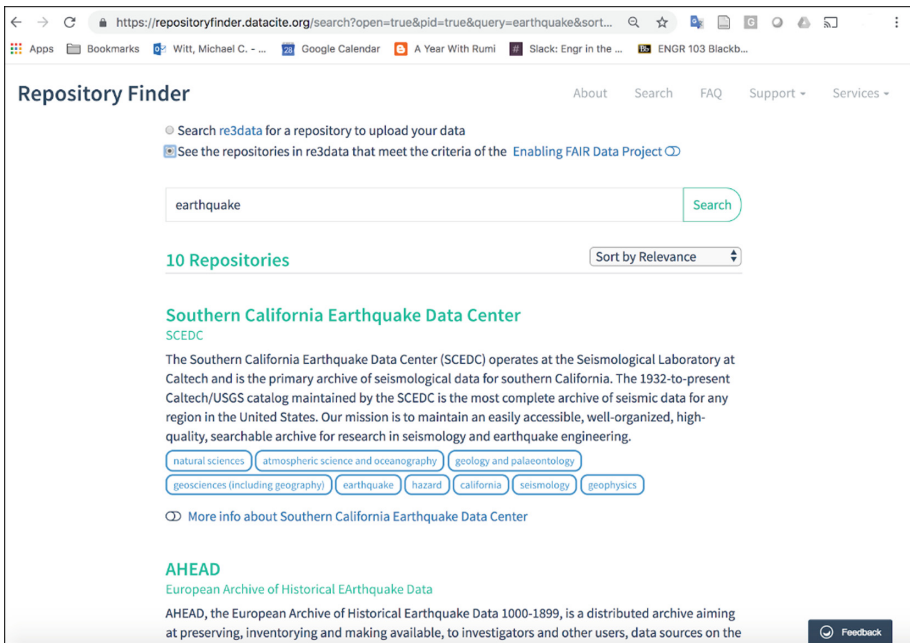


Fig. 4. Search results in the repository finder.

2.4 User Testing

In between the third and fourth development sprints, a prototype of the tool was made available for the purpose of user testing, which occurred in three, separate studies. The first two studies were held at the 2018 ESIP Summer Meeting to leverage the diversity of the attendees’ data roles and responsibilities. The first study that was conducted by

the ESIP Usability Cluster²⁶ engaged fifteen users in a session with a focus on usability that combined the focus group and user study techniques so that the users could provide feedback regarding specific, task-based interactions with the tool in a moderated fashion. The second study was conducted by the ESIP Usability Cluster Chair and Fellow using the one-on-one, in-person user study technique with a total of five users who fit the general “researcher” persona. During the user study session, each subject was first asked to tell a “user story” and describe a goal they would like to accomplish when looking for a repository. The subject was then asked to use the tool to perform specific, defined tasks based on the “user story” scenario and the goal. The subject was asked to think aloud while performing the tasks, so that the subject could share the thought process as they determine how to use the tool to accomplish the tasks. Feedback from users identified several concerns that aligned with usability issues that are outlined and discussed in Jakob Nielsen’s 10 Usability Heuristics for User Interface Design [10]. In particular, users indicated that improvement in the following areas could have significant impact on the user interface and experience of the tool: the identity and messaging of the tool (e.g. what can the tool do, and why would I use it?), the authenticity and understandability of the content (e.g., institutional repositories not included in results, too many filters that only repository managers would understand, and lack of information regarding how the repositories’ records are curated and can be updated in re3data); and the overall design approach (e.g. be more user-centric by providing visibility of system status, minimizing use of jargons, adding support documentation such as an FAQ, and implementing user friendly aesthetics for the user interface).

In the third study, a commercial firm was hired to interview users while guiding them through use of the tool. Subjects were recruited by project team members and included twenty-four individuals who identified themselves as either domain researchers or repository “champions” such as librarians, curators, data facility staff, or other data professionals. Test subjects described their backgrounds coming from Astronomy, Astrophysics, Atmospheric Sciences, Ecology, Environmental Science, Geodesy, Geography, Geoinformatics, Geology, Geomorphology, Geophysics, Hydrogeology, Hydrology, Oceanography, Paleooceanography, Palaeontology, and Polar Science. All subject classifications from the selected DFG subject areas were represented by at least one user with the exception of 316-01 Geochemistry, Mineralogy, and Crystallography. Online web sessions were conducted and recorded using Zoom with each session lasting between twenty and thirty minutes each. An interviewer gave a brief introduction to the tool and its purpose before guiding the user through the task of searching for a repository to deposit their data. Users were asked if the terminology and prompts were clearly understood as well as to evaluate the results for accuracy and relevance. Were there any repositories that were not relevant in the result list, and were the top results the most relevant? Were there any repositories they would expect to see in the results that were not there? Lastly, interviewers asked if the user’s expectations were met by the tool and to share any additional input, ideas, or improvements that could be incorporated in the final development cycle of the tool or

²⁶ <http://wiki.esipfed.org/index.php/Usability>.

potential future development. Recordings of the sessions and interview notes were analyzed and a summary report reaffirmed users' perception of the need for such a tool and suggested improvements needed to the utilization of space and level of detail presented in the user interface, the mechanism for searching, and the metadata quality and completeness of the information about repositories in re3data. The user interface was refined to give collapsed results that can be expanded to show more detail, and the original hierarchical drop-down list of subjects was replaced with an auto-completing keyword search and a simple link to display a list of repositories that meet the criteria of the Enabling FAIR Data community initiative. The quality of the results given by the tool are impacted directly by the quality and completeness of metadata describing each repository in the re3data registry. To begin to address concerns with metadata, the re3data editorial board met for a two-day workshop hosted by the Karlsruhe Institute of Technology to focus on enhancing and bringing more consistency to the records for repositories in the Earth, space, and environmental sciences. They collectively edited a representative sample of records and compared notes to discuss issues and build consensus around editorial practices, in particular, for the repository attributes that impact Repository Finder and other tools like it in the future that may be built using re3data's API. The most accurate information about a repository can be provided to re3data by the repository managers themselves; to encourage their participation, a one-page guidance document was prepared and disseminated to the community by AGU. Users of Repository Finder and re3data continue to submit enhancements and corrections that refine and improve the quality of the registry and the results provided by the tool.

3 Challenges and Future Work

The project faced familiar constraints of a short timeline and limited resources; while we were able to engage a significant cross-section of stakeholders who work with data in the context of scholarship in Earth, space, and environmental sciences, it was not representative of the domain as a whole. Input was not statistically significant and relied on convenience samples of parties engaged in the initiative with a strong representation bias from North America and Europe. More time and broader engagement would result in a more robust consensus. The decision tree, interviews, and user tests were intended to inform the development of the tool in a practical and direct manner; they were not designed to stand alone as formal research studies. Problems related to metadata extended beyond the completeness of records in re3data: differences in terminology and the lack of widely adopted controlled vocabularies for the domain and subdomains as well as data types raised concerns in user testing and limited the potential functionality of the tool, e.g., the ability to search repositories by the type of data they accept. In particular, the DFG subject classification was limiting, for example, by not including commonly used subject terms such as "environmental science". Within the re3data schema, the most flexibility to overcome this limit exists in adding a variety of different keywords that represent the same subjects with different names and in the description, in particular, to add very specific subdomain terminology, instrumentation, and data formats. There is an inverse relationship between the number of

attributes that are used to filter searches and the quantity of results those searches will produce. In terms of promoting FAIR, many of the practices that were reported by repositories were either not widely adopted, or in some cases, not cataloged in the registry. For example, an early iteration of the tool limited results to only certified repositories, but this excluded too many repositories that researchers recognized as being relevant to their research, and in some cases, yielded no results at all. There is also an important role for repositories and data facilities that are provided locally by institutions; however, the tool does not know the affiliation of the user, so it was not possible to include relevant institutional repositories in the results. The current version of Repository Finder was limited in scope to the use case of a researcher selecting a repository to deposit their data, but discussions with other TAGs and from user testing suggested many other use cases that could be motivated by publishers, journals, funders, societies, and other drivers that could be explored. Other potential future work includes updating the criteria to reflect the metrics coming out of GO FAIR and other, related initiatives as well as extending this approach to other domains outside of the Earth, space, and environmental sciences and to other lists of recommendations that may exist or emerge in the future.

Acknowledgements. Enabling FAIR Data was made possible by a grant from the Laura and John Arnold Foundation to the American Geophysical Union; program manager, Shelley Stall. Sabbatical support for Michael Witt was provided in part by the Pufendorf Institute of Advanced Studies at Lund University.

References

1. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016)
2. Cruse, P., Servilla, M.: Workflow Recommendations for Enabling FAIR Data in the Earth, Space, and Environmental Sciences, Zenodo (2018). <http://doi.org/10.5281/zenodo.1445839>
3. Buddenbohm, S., De Jong, M., Priddy, M., Moranville, Y., Ribbe, P.: Open Data in the Humanities Platform: Humanities at Scale: Evolving the DARIAH ERIC, DARIAH; DANS-KNAW (2017). <https://hal.archives-ouvertes.fr/hal-01686320>
4. ISO 16363:2012 - Space data and information transfer systems – Audit and certification of trustworthy digital repositories (2012). <https://www.iso.org/standard/56510.html>
5. Core Trustworthy Data Repositories Extended Guidance (2018). <https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf>
6. Witt, M.: Co-designing, co-developing, and co-implementing an institutional data repository service. *J. Libr. Adm.* **52**(2), 172–188 (2012)
7. Whyte, A.: Where to Keep Research Data: DCC Checklist for Evaluating Data Repositories Version 1.1, Digital Curation Centre (2016). <http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data/where-keep-research-data>
8. Plante, R., Witt, M.: Interview Questions for Determining Data Repository FAIR Compliance, Zenodo (2018). <http://doi.org/10.5281/zenodo.1432515>
9. Rücknagel, J., et al.: Metadata Schema for the Description of Research Data Repositories: version 3.0, p. 29 (2015). <http://doi.org/10.2312/re3.008>
10. Nielsen, J.: 10 Heuristics for User Interface Design, Nielsen Norman Group (1995). <https://www.nngroup.com/articles/ten-usability-heuristics>