Ecological Informatics 70 (2022) 101729

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Asking nicely: Best practices for requesting data

Stephen R. Midway[a,*], Nicholas A. Sievert[b], Abigail J. Lynch[c], Joanna B. Whittier[d], Kevin L. Pope[e]

[a] Louisiana State University, 2257 Energy, Coast & Environment Building, Department of Oceanography and Coastal Sciences, Baton Rouge, LA 70803, United States of America
[b] Missouri Cooperative Fish and Wildlife Research Unit, School of Natural Resources, University of Missouri, Columbia, MO, United States of America
[c] U.S. Geological Survey, National Climate Adaptation Science Center, 12201 Sunrise Valley Drive, Reston, VA, United States of America
[d] School of Natural Resources, University of Missouri, Columbia, MO, United States of America
[e] U.S. Geological Survey—Nebraska Cooperative Fish and Wildlife Research Unit, School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, United States of America

ARTICLE INFO

Keywords:
Data provider
Data requester
Database
Agency

ABSTRACT

Compiling disparate datasets into publicly available composite databases helps natural resource communities explore ecological trends and effectively manage across spatiotemporal scales. Though some studies have reported on the database construction phase, fewer have evaluated the data acquisition and distribution process. To facilitate future data sharing collaborations, Louisiana State University surveyed data providers and requestors to understand the characteristics of effective data requests and sharing. Data providers were largely U.S. natural resource agency personnel, and they reported that unclear data requests, privacy issues, and rigid timelines and formats were the greatest barriers toward providing data, but that they were motivated by improving science and collaboration. Data requestors identified challenges such as evolving needs, standardization issues, and insufficient resources (time and funding) as barriers to compiling data for these types of efforts. In a time of big data, open access, and collaboration, significant scientific advances can be made with effective requests and inclusion of data sets into larger and more powerful databases.

1. Introduction

By many accounts, ecology is in a time of big data (Farley et al., 2018; Hampton et al., 2013; Peters et al., 2014), which has seen the emergence of databases created from many datasets of varying spatiotemporal coverage. Individual datasets typically hold great value to their creator and target user; even so, datasets combined into larger databases can provide additional value. Composite datasets provide utility to a wider user group, allowing for important new discoveries (Aubin et al., 2020; Heberling et al., 2021; Whittier et al., 2016) and emergent findings (e.g., cross-scale interactions; Soranno et al., 2014) not available from a single dataset. Ranging from DNA (GenBank; NCBI, 2016) to the ecosystems (National Land Cover Database; Dewitz, 2019), data are increasingly available for different ecological scales and biota. Regional, continental, and global ecological challenges are likely only to be met with databases of commensurate scale; therefore, identifying best practices for acquisition and transfer of individual datasets can help reduce barriers.

Valuable ecological datasets are created every day in diverse ways. However, datasets consistently collected and maintained by governmental natural resource agencies (e.g., fish and wildlife agency, department of water quality) hold particular promise due to their surveillance design (Nichols and Williams, 2006), often large spatiotemporal scale, and dedicated resources for curation, storage, and other stewarding. Agency datasets are further enhanced by the ubiquity of computers, which has enabled rapid and comprehensive data entry, quality assurance and quality control, safety for long-term storage, improved preparation for statistical software, and quick transfer of data to different users. In the field of natural resources, a number of databases illustrate the successful cohesion of composite datasets. For example, the LAGOS-US database provides users with unprecedented information about U.S. lakes (Cheruvelil et al., 2021), FishBase is a popular global information system on fishes (Froese and Pauly, 2021) and CreelCat (Lynch and 16 co-authors., 2021) is a new database entirely generated from U.S. state agency contributions. Such databases are fueling a change in the scales at which ecological questions can be addressed

* Corresponding author.
E-mail address: smidway@lsu.edu (S.R. Midway).

https://doi.org/10.1016/j.ecoinf.2022.101729
Received 18 February 2022; Received in revised form 18 June 2022; Accepted 18 June 2022
Available online 21 June 2022

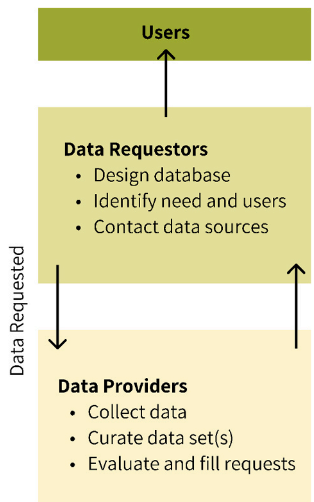

Fig. 1. Conceptual diagram of roles in the data acquisition process. Typically, data requestors have designed a database and need to contact data providers to place a request. Data providers are the curators of the data and evaluate the data requests for fulfillment. Users are those that may benefit from the product developed by the data requestors, and users may include the requestors, the providers (i.e., the agency), or external groups or individuals.

(McManamay and Utz, 2014).

Although the creation of a useful ecological database may seem like a straightforward task, it is often a complicated process for which few blueprints exist (Midway et al., 2016; but see Soranno et al., 2015a). In general, creating a database can be decomposed into two phases: 1) the data acquisition phase and 2) the database construction phase. The data acquisition phase includes everything from identifying and evaluating the idea and need for the database, locating candidate datasets for inclusion, navigating data sharing ethics, transferring or acquiring data, and developing metadata or other data descriptors that allow users to fully understand what is included in a given data set. (But note that in our context the data acquisition phase does not include collection of raw data in the field.) Often, the data acquisition phase requires a lot of interpersonal communication between different individuals and groups as data requestors make a case for their need and data providers evaluate and (hopefully) fill the requests (Fig. 1). The database construction phase is somewhat more autonomous for the internal database team, as they meet the challenges of creating standards that will harmonize different datasets that may have intrinsically different observation levels or units, working to enhance database longevity, providing access to users, creating an accessible database that is not complicated, and marketing the database so that potential users know of its existence. Despite this autonomy, there are still interpersonal and team dynamics that can greatly improve project outcomes (Cheruvelil et al., 2014; Cheruvelil and Soranno, 2018).

All steps of database creation warrant new and continued study so that useful databases can be created with minimal friction. Several studies have investigated the database construction phase (Kolb et al., 2013; McLaughlin et al., 2001; Soranno et al., 2015a), but very few have examined the data acquisition phase. Here, we define the data acquisition phase as the actions associated with external users (i.e., those outside the agency) acquiring the data from the agency that collected it. Although the acquisition of data from a public agency may seem like a straightforward transaction—and it often can be—there are characteristics of data requests that may increase or decrease the probability of having that request fulfilled along with affecting the quality of information that comes with a data request. The goal of this study was to identify best practices and make recommendations about the data acquisition process. Using two separate and distinct survey methods, we analyzed feedback from data requestors (i.e., database creators) as well as data providers (i.e., dataset curators) to better understand their motivations and what they perceive to be the characteristics of a successful data request and transfer. More effective data requests may lead to higher fulfillment rates, faster responses, and even higher quality data being provided, and improvements in data acquisition should lead to better outcomes in database creation, quality, and continuity.

2. Materials and methods

To fully understand the external data request process (hereafter, for the sake of simplicity, referred to as a *data request*), we studied both data providers and data requestors because each group plays an important role. We selected different methods for working with and evaluating each group. To best understand the data providers, Louisiana State University (LSU) developed an anonymous (online) survey, because a survey can be quickly distributed to many participants, and we wanted to invite multiple data providers from any U.S. state agency willing to participate. For data requestors, we opted to use a focus group composed of data requestors who had developed composite databases that are currently being widely utilized. A focus group also allowed for a more flexible environment to share potentially unique thoughts and experiences (that a standardized survey may not capture).

2.1. Data providers: online survey

During the summer of 2021, LSU developed a survey to document 1) the attributes of data providers (with questions about how long they have held their job, how many data requests they typically complete, and any relevant training), 2) how data providers perceive their agency's capacity to fulfill data requests, and 3) what the real or perceived barriers and benefits are in fulfilling data requests. The survey was designed to be a short (5–10 min to complete), all questions were optional, and prior to releasing, the survey was approved by the Institutional Review Board of Louisiana State University (IRBAM-21-0722). The approved survey is provided (Appendix 1).

Analysis of the survey data was primarily descriptive, except for the ranked barriers and benefits, which were evaluated with linear models. Specifically, participants identified and then ranked the barriers and benefits (separately) toward meeting data requests (Table 1). Ranks ranged from 1 (top rank) to 7 for barriers and to 6 for benefits, although not all factors were ranked by each participant and therefore some rankings were missing. Any unranked factors were assigned the bottom rank (i.e., 7 for barriers; 6 for benefits), which reflects participant experiences. These last-rank assignments also serve to avoid the biases associated with excluding unranked factors in analyses; for example, if a specific factor was often excluded, it could still have a relatively high mean rank based on the times it was ranked and if the exclusions were not accounted. After all factors had values, a beta regression (Ferrari and Cribari-Neto, 2004) was used to analyze their means and test for differences. Rankings are a form of interval data; thus rankings were simply transformed by dividing each rank by the maximum possible rank, and then subtracting 0.01 to avoid a value of 1 because a beta-distributed interval variable must occur between 0 and 1. The beta regression was

Table 1
Possible barriers and benefits data providers may identify when evaluating an external data request, and which survey takers were asked to rank.

Barriers	Benefits
Accessing data	Agency goodwill
Formatting data	Collaboration
Not a priority	External access
Privacy issues	Improve science
Time for request	Interactive tools
Transmitting data	Public access
Unclear request	

essentially a beta ANOVA because the independent variables were categorical factors with the interval-transformed rankings as the response variable. After the beta regression was run, estimated marginal means were computed (Lenth, 2021) and pairwise comparisons using a Tukey post hoc adjustment (Midway et al., 2020) were carried out to identify which specific barriers (and benefits) statistically differed from each other.

For the questions about what constitutes a good data request, we opted to include a text response to allow us to identify and enumerate common terms. To visualize the results of text-based responses about good and bad data requests, we used a word cloud per request category, in which the font size is proportional to the frequency of the term based on all responses. We excluded words that 1) were common articles (e.g., *the*, *an*), 2) had a frequency of 1, and 3) used any version of the terms "data" and "request" because these terms were in the question and were often repeated in the response.

The final survey question asked participants to identify the direction and magnitude of influence (ranging from -10 for negative influence to $+10$ for positive influence with 0 indicating no influence) for different potential factors influencing a data request. Factors of a data request include *beneficiary*, *deadline*, *format*, *frequency of request*, and *workflow*. In the survey, each factor was split into two subfactors for ranking and comparison; e.g., the *deadline* factor asked separately about a rigid deadline and a flexible deadline, and the workflow factor asked separately about a direct data request and an indirect data request (Appendix 1). Similar to analysis of barriers and benefits, a beta regression was run on the data due to its interval nature. However, we did not run one model with all factors in it. Rather, we ran five beta *t*-tests that allowed us to evaluate which factor, because the factors were presented as two subfactors, was significant as indicated by the subfactors being statistically different ($\alpha = 0.05$) in their importance. Post hoc adjustments were not made because each *t*-test included only one possible means comparison that was coherent with the outcome of the test.

2.2. Data requestors: focus group

We assembled a focus group to explore the motivations of database creators as well as the challenges and opportunities of creating composite natural resources databases. We invited ten participants stratified across academic, agency, and nongovernmental organization backgrounds, along with considerations for gender and professional experience; these invitations were strategic to engage with individuals with a particular expertise across a diverse range of experience, employment, and perspective. As five invitations were declined (based on logistics), our focus group included five participants with experience creating composite databases for natural resources data in both terrestrial and aquatic systems. SRM and AJL served as moderators and NAS served as note taker during the focus-group discussion, and the discussion was recorded (with consent from all participants). The following prompts were used to facilitate discussion among the focus-group participants:

1. What motivated you to undertake developing a composite database from multiple disparate datasets?
2. What do you see as being the major challenges to developing a composite database?
3. What have you learned from developing a composite database?

The focus group was designed to take place over approximately 2 h (including breaks), all topics were optional, and prior to convening, the agenda was approved by the Institutional Review Board of Louisiana State University (IRBAM-21-0722). Notes on the discussion from the focus group were summarized to identify key points.

3. Results

3.1. Data providers: online survey

3.1.1. Description of survey participants

A total of 45 agency personnel completed the survey, which resulted in a 56% participation rate (based on $n = 80$ survey invitations). Participants represented 25 U.S. states (Fig. 2) and reported a mean of 13 years of their career providing data, with a range from 2 to 40 years. The majority (60%) of participants reported having not received formal training (classes, workshops, etc.) in data management whereas the remaining 40% indicated some formal training. The number of formal

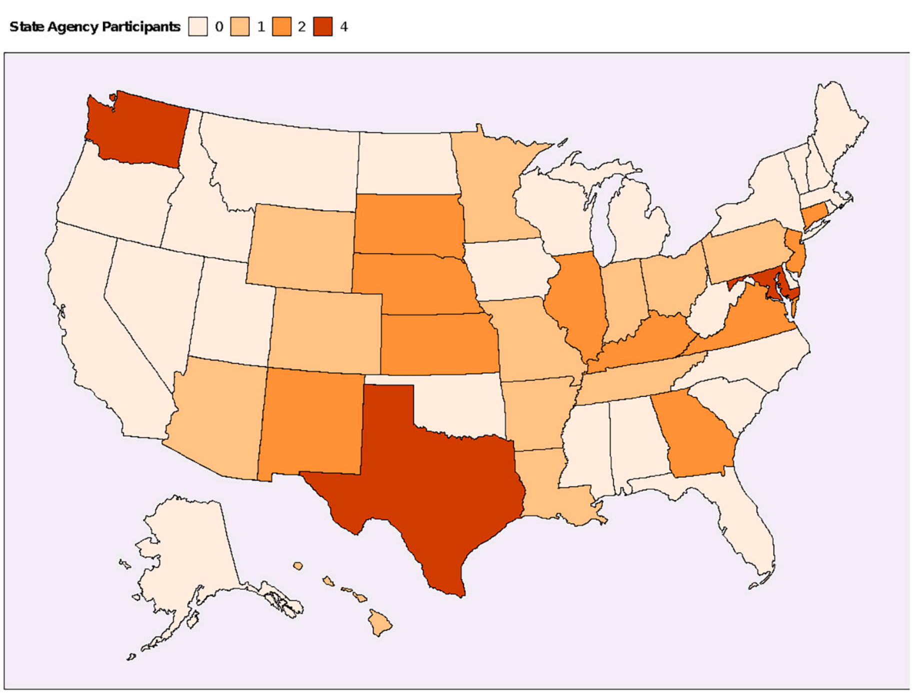

Fig. 2. Map of U.S. states with shading to indicate the number out of the 45 agency data providers who responded to the survey (out of a total of 80 invitations).

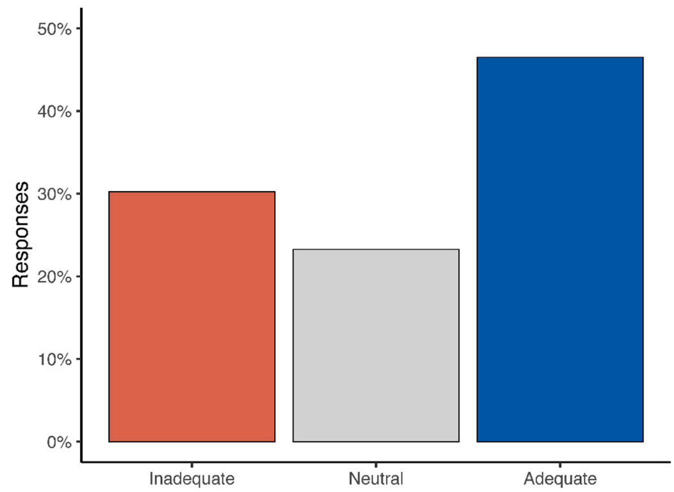

Fig. 3. Categorized responses to "How adequate do you think your agency's data request system is for completing external data requests?" Respondents selected one of three categorized answers.

data requests that participants completed in a typical year was highly variable with a range from 3 to 1200. Only four responses to this question were >100 (suggesting many small data requests) and the median number of annual data requests was 12. Finally, 63% of participants reported that their agency has some public data portal (e.g., website) where some amount of data can be queried and downloaded.

3.1.2. Consideration of data requests

When asked about their agency's (internal) system for handling data requests, the most common response (16 of 43 responses) was *Somewhat Adequate* with another four reporting the system to be *Adequate* (Fig. 3, but which reports net adequacy). Net inadequacy was reported by 13 people and 10 thought of their agency's system as neutral in its ability to respond to data requests.

Most participants (32 out of 43) reported that their agency gave *some* consideration to data requests in the data management workflow (Fig. 4). When asked at a finer resolution about the project phases that may consider data requests, the respondents reported the lowest data request integration during the data collection phase and the greatest integration at the data storage and archiving phase (Fig. 4).

3.1.3. Barriers and benefits

The highest ranked barrier to completing data requests was *Unclear Request*, followed by *Privacy* and *Time* (Fig. 5A). The highest ranked benefit to completing data requests was *Improve Science*, followed by *Collaboration* and *Public Access* (Fig. 5B). For analysis of both barriers and benefits, poorly ranked factors were heavily weighted by the assigned bottom rankings, meaning these factors were not identified by data providers as being relevant enough to include. Pairwise comparisons found some differences ($p < 0.05$; Fig. 5) in mean rankings, although most differences were between the extreme rankings; in other words, the top ranked barrier or benefit was not statistically different from other highly ranked barriers or benefits.

3.1.4. Word clouds

In response to the survey question asking for a text response describing the characteristics of a good data request the most frequently used terms were *clear* (19 times), *specific* (13 times), *good* (6 times), and *concise* (5 times; Fig. 6). All other words were used 4 or fewer times. In response to the survey question asking for a text response describing the characteristics of a bad data request the most frequently used terms were *vague* (11 times), *unclear* (8 times), *broad* (7 times), and *just* (5 times; Fig. 6). All other words were used 4 or fewer times.

3.1.5. Influences on completing a data request

Of the five factors we investigated (*beneficiary*, *deadline*, *format*, *frequency of request*, and *workflow*), *deadline* and *format* were significant effects based on the significant difference between their subfactors (Fig. 7). A flexible deadline was significantly preferred over a rigid deadline ($p = 0.004$), and a flexible format was significantly preferred over a rigid format ($p < 0.001$). Differences in means were calculated for

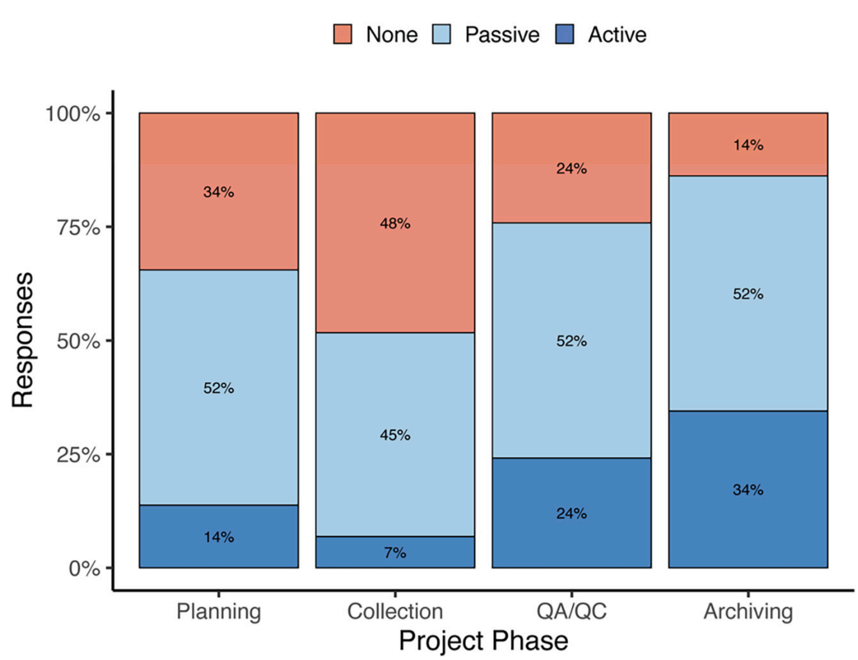

Fig. 4. Categorized responses to "How are external data requests integrated into your agency's data management workflow?" Respondents selected one of three categorized answers (*none*, *passive*, *active*) for four phases (*planning*, *collecting*, *quality assurance, and quality control* [QA/QC], *achieving*).

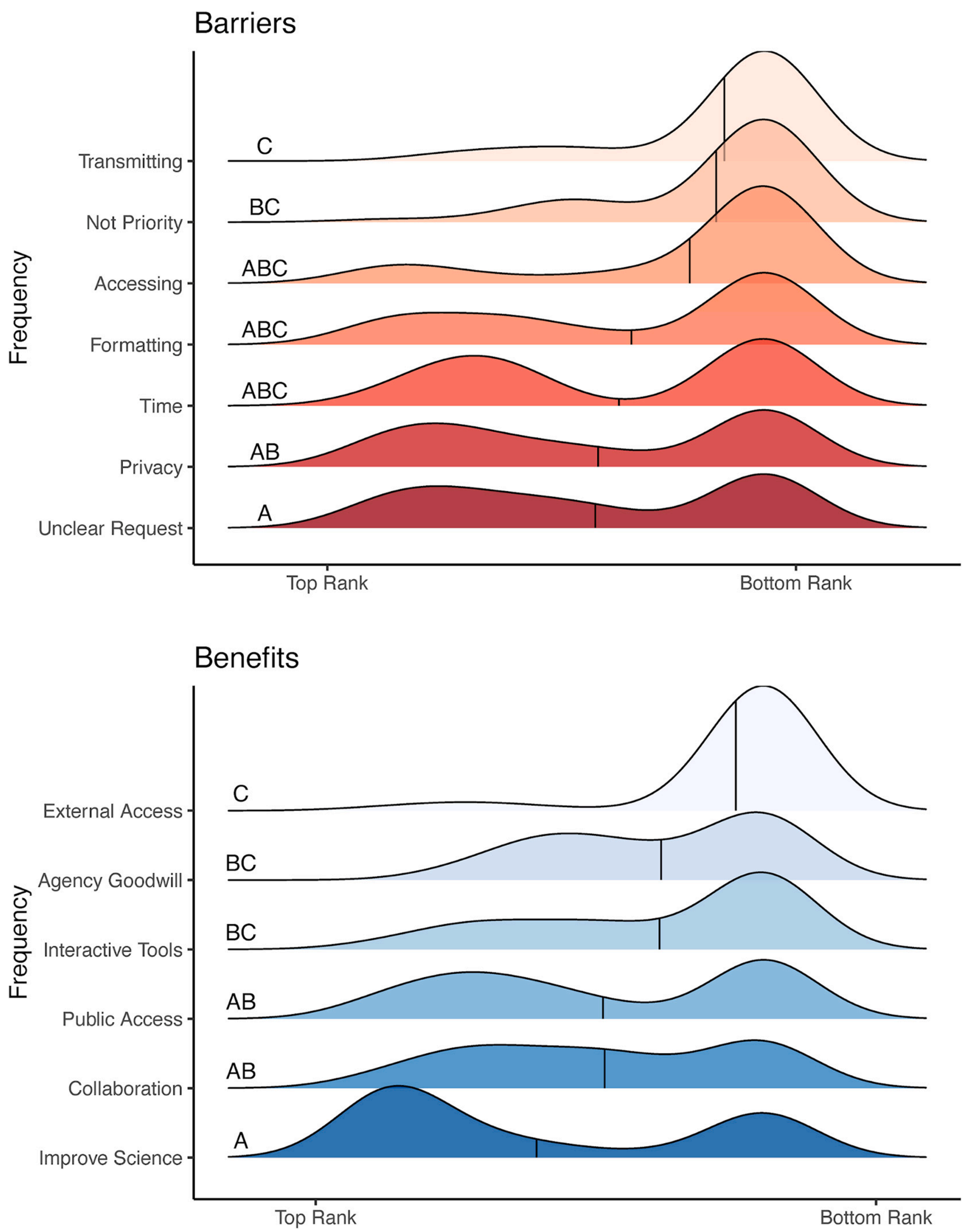

Fig. 5. Barriers and benefits, as reported by agency data providers, associated with filling a data request. For both barriers and benefits, the top mean ranked factor is at the bottom of the respective panel and each factor's density represents the rankings for that individual factor. Letters to the left of the density panels are group assignments based on a Tukey post hoc comparison, where factors not sharing letters are significantly different from each other. Vertical lines in the density features represent the mean rank for that factor. The greatest color tint is provided to the top ranked barrier or benefit with the lowest tints reflecting the lowest ranked barriers and benefits.

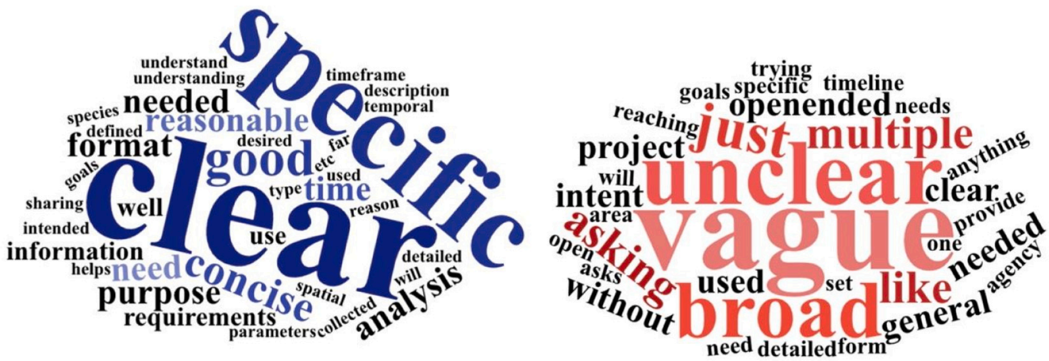

Fig. 6. Word clouds representing terms associated with good data requests (left) and bad data requests (right). The size of the words corresponds with the frequency of the term data providers used in text-based survey question responses.

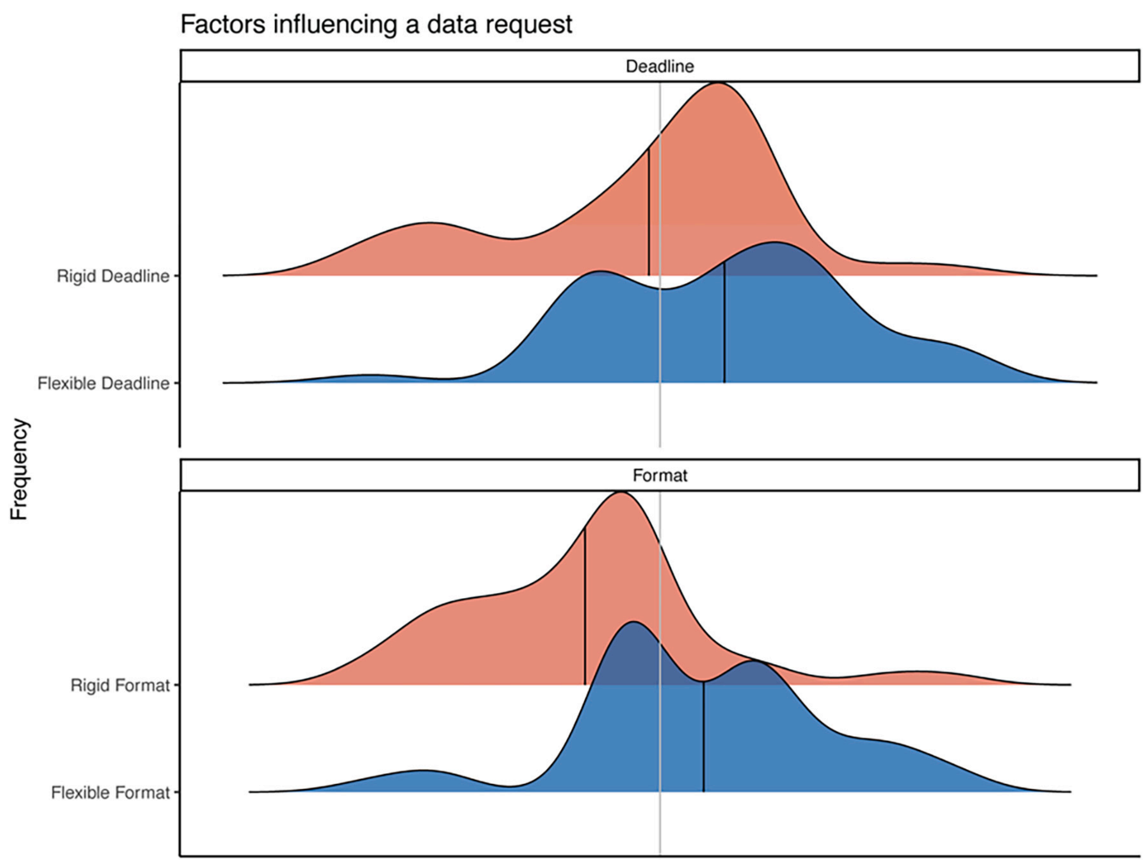

Fig. 7. Density plots comparing the subfactors for deadline and format, which were the only two factors reporting a significant difference. In both cases, the flexible response was significantly favored ($p < 0.05$) over the rigid response. Color is used only to differentiate the density features, the gray vertical lines represent 0, and the black vertical lines represent the mean responses.

the factors of beneficiary, repeated request, and workflow; however, the subfactors were not found to be statistically significant from each other.

3.2. Data requestors: focus group

3.2.1. Description of survey participants

Five individuals participated in the focus group, which was held virtually on August 13, 2021. These individuals were all currently working in the natural resources profession, had prior experience with compiling databases, and comprised 60% women and 20% persons of color. Further, 60% of the participants were employed in academia and the remaining 40% were employed with a non-governmental organization.

3.2.2. Motivation for database creation

Participants shared a wide variety of motivations for developing composite databases, including filling needs for both personal research projects and employer objectives. Participants noted that in some cases motivation for development of these databases occurred after initial efforts were made to compile data for a specific research or management need, once the potential values of the data to partners, employers, or others beyond the initial data user were identified. In other instances, motivation to create a composite database came from a need to enhance existing efforts in which data being used to support research or management efforts existed in an unorganized structure. Participants noted that oftentimes their efforts to develop composite databases came about later in the research process as the value of data being initially collected for a short-term endeavor was recognized, which prompted efforts to create a more comprehensive and accessible database.

3.2.3. Challenges

A variety of challenges in database development were identified by the participants. The issue of standardization for merging data from multiple sources was identified by several participants as the most challenging issue faced in their efforts. Despite inadequate metadata being a long recognized issue related to data management (Michener, 2015), the frequent lack of suitable metadata or information required to interpret data was another key issue identified by participants. Adequate resources to support and sustain the development and maintenance of these databases was also a challenge that was identified by many participants. Participants noted that the amount of time required for these efforts is substantial, and either limited funding or limited work hours when factoring in other responsibilities can be a barrier for developers to create and maintain such databases. Another important barrier for developers, in some instances, is data-security issues, especially centered on data ownership and multiple and varied agreements for data sharing and use.

3.2.4. Lessons learned

Participants identified several important lessons learned. There is great value in planning for the long term and collaborating with likely data providers and users about design and management of the database. Similarly, recognizing that no matter how much planning is involved, the work always takes longer and requires more time and effort than anticipated. It can be especially challenging to get people to share or contribute data to a database out of fear of misuse or creating additional, undue burdens. As such, relationships with potential data providers must be fostered to build trust and support. There is value in making data public—open science is a worthwhile cause (Reichmann et al., 2011; Soranno et al., 2015b)—though one must understand privacy issues and incorporate concessions when prudent. Finally, be willing to share, without compromising the trust and support of the data providers; as one participant noted “the coolest analyses that will be done using the database will likely be done by someone other than the database creator.”

4. Discussion

Large composite databases hold the promise of addressing major ecological challenges (Whittier et al., 2016) and improving the data acquisition process for databases should help create better and more useful databases. Challenges begin as early as the planning stages (Midway et al., 2016), where ideas, data availability, and logistics all need to be balanced. For instance, all databases require some design and planning; however, database creation often occurs after initial efforts have been made for a smaller project and the potential value of the database only then comes into focus. Obviously, database design is an iterative process that may include trial and error; however, it remains important to have a central vision and plan for the data that needs to be acquired and have good reasons that specific data need to be included. A lack of clarity and understanding the database vision may create challenges in providing a clear data request to data providers, who often appreciate knowing the intended use and benefit of their data. Despite these initial challenges data requestors should know that many data providers reported an adequate ability to handle data requests and that consideration of data requests are typically integrated in all phases of data handling (from planning to archiving).

Assuming the data provider has the appropriate data for a request there are several things that a data requestor can do to improve the quality of their request and the subsequent outcome of the request being fulfilled. The top barrier to completing data requests is an unclear request. We cannot provide specific universal guidance on a clear data request because requested data are too variable to generalize. However, a clear request not only outlines the specifics of the desired data but should demonstrate knowledge of the sampling program and agency handling of the data. Program and agency knowledge generally can be obtained by reading online information about a program and agency. Spending time to review program reports will help the data requestor identify the agency names for specific variables of interest as well as the period of collected datasets. In addition, learning an agency's mission and priorities will help a requestor understand the legal limitations placed on an agency, as well as how the requestor's work could benefit the agency.

Providing clarity in the intended data application, the variables of interest, the spatiotemporal domain, the preferred data format, and the request timing are all important components of a clear data request (see Boxes 1 and 2 for examples). It is also important to understand what the perceived benefits are to the data provider. Although data sources may be compelled to provide data by law or public access regulations, it remains important to understand that the individuals fulfilling the data requests often appreciate understanding how a data request benefits them, their agency, or the larger project. For example, data providers told us that their top ranked benefit toward filling a data request is to improve science, and the benefits of collaboration and public access were also important.

The increased digitization and automation of data transfer is undoubtedly a positive development in the ability to share data. Yet, with all this improved efficiency and electronic infrastructure, it remains important to understand that there are still people who put a lot of time and energy into collecting and curating data, and even when those data are intended for public access, a clear, concise, and friendly request is the best way to demonstrate respect and appreciation to the data

Box 1
Example of a poor data request. Note that characteristics of a poor data request include a lack of clarity about the specific data, few (or no) research objectives, short or ambiguous deadline, lack of any formatting requirements, and overall vagueness.

Hi Tom,

Can you send me all your unicorn data? I don't have a specific research topic just yet—I want to take a look at these data and see what I might find. You can just send over the whole database. I'll reach out if I have any questions.

I've got a meeting with the American Unicorn Society next week, so I need this request ASAP.

Appreciate the help!

Craig P.

Box 2
Example of an effective data request. Note that characteristics of an effective data request include clear and concise research objectives and data descriptions, flexible and reasonable deadline(s), data provider acknowledgment in future products, and guidance on data formatting.

Hi Tom,

I'm interested in examining if recruitment campaigns have influenced the number of unicorn hunters across southeastern states. I would like to compare unicorn license numbers pre and post recruitment campaign, and between counties targeted by the campaign and those that were not.

*Could you share with me the **2019** and **2021 unicorn license sales numbers by county** for all counties in your state?*

*Might it be possible to fulfill this request **within one month**? The preferred format would be variables in columns and observations in rows, all within a CSV file, but I am flexible to other formats if something else is easier for you.*

I will combine data across states, creating a composite database, for a complete analysis. I will be presenting to the American Unicorn Society in two months and would like to be able to run some correlation analyses before that meeting. I will acknowledge you and your agency as one source for these data in my presentation. Additionally, I plan to publish these findings in a peer-reviewed journal and will acknowledge your data contributions in my manuscript.

If there is any additional information I can provide to assist with this request, please let me know.

I appreciate your help with this specific request and thank you for managing such a valuable database.

Sincerely,

Joanna W.

provider. Although it remains speculation, some unfulfilled data requests are likely the result of fear of data misuse, misrepresenting results, or other unknowns. Keeping an open dialogue—through a combination of email, phone calls, and in-person communication—is a great way to develop relationships between data requestors and data providers and ultimately diminish or eliminate the unknowns that may hamper data requests. Database creators and data sources likely share the same principles and desired outcomes: to make the best available data into good science that can improve natural resource and ecological outcomes. In this way, data providers and data requesters are on the same page and through improved communication in the data acquisition process good outcomes should lead to better and more powerful databases.

Declaration of Competing Interest

None.

Data availability

The data that has been used is confidential.

Acknowledgements

We thank the participants who completed the survey and the participants of the focus group. Surveys described herein were organized and implemented by Louisiana State University and were not conducted on behalf of the U.S. Geological Survey. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The Nebraska Cooperative Fish and Wildlife Research Unit is jointly supported by a cooperative agreement among the U.S. Geological Survey, the Nebraska Game and Parks Commission, the University of Nebraska, the U.S. Fish and Wildlife Service, and the Wildlife Management Institute.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2022.101729.

References

Aubin, I., Cardou, F., Boisvert-Marsh, L., Garnier, E., Strukelj, M., Munson, A.D., 2020. Managing data locally to answer questions globally: the role of collaborative science in ecology. J. Veg. Sci. 31 (3), 509–517.
Cheruvelil, K.S., Soranno, P.A., 2018. Data-intensive ecological research is catalyzed by open science and team science. BioScience 68 (10), 813–822.
Cheruvelil, K.S., Soranno, P.A., Weathers, K.C., Hanson, P.C., Goring, S.J., Filstrup, C.T., Read, E.K., 2014. Creating and maintaining high-performing collaborative research teams: the importance of diversity and interpersonal skills. Front. Ecol. Environ. 12 (1), 31–38.
Cheruvelil, K.S., Soranno, P.A., McCullough, I.M., Webster, K.E., Rodriguez, L.K., Smith, N.J., 2021. LAGOS-US LOCUS v1.0: data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous US. Limnol. Oceanogr. Lett. 6 (5), 270–292.
Dewitz, J., 2019. National Land Cover Database (NLCD) 2016 Products: U.S. Geological Survey Data Release. https://doi.org/10.5066/P96HHBIE.
Farley, S.S., Dawson, A., Goring, S.J., Williams, J.W., 2018. Situating ecology as a big-data science: current advances, challenges, and solutions. BioScience 68 (8), 563–576.
Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. J. Appl. Stat. 31 (7), 799–815.
Froese, R., Pauly, D., 2021. FishBase. World Wide Web electronic publication.
Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A. L., Duke, C.S., Porter, J.H., 2013. Big data and the future of ecology. Front. Ecol. Environ. 11 (3), 156–162.
Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. Proc. Natl. Acad. Sci. 118 (6).
Kolb, T.L., Blukacz-Richards, E.A., Muir, A.M., Claramunt, R.M., Koops, M.A., Taylor, W. W., Sutton, T.M., Arts, M.T., Bissel, E., 2013. How to manage data to enhance their potential for synthesis, preservation, sharing, and reuse—a Great Lakes case study. Fisheries 38 (2), 52–64.
Lenth, R.V., 2021. emmeans: Estimated Marginal Means, aka Least-Squares Means. R Package Version 1.7.0.. https://CRAN.R-project.org/package=emmeans.
Lynch, A.J., 16 co-authors., 2021. The US inland creel and angler survey catalog (CreelCat): development, applications, and opportunities. Fisheries 46 (11), 574–583.
McLaughlin, R.L., Carl, L.M., Middel, T., Ross, M., Noakes, D.L., Hayes, D.B., Baylis, J.R., 2001. Potentials and pitfalls of integrating data from diverse sources: lessons from a historical database for Great Lakes stream fishes. Fisheries 26 (7), 14–23.
McManamay, R.A., Utz, R.M., 2014. Open-access databases as unprecedented resources and drivers of cultural change in fisheries science. Fisheries 39 (9), 417–425.
Michener, W.K., 2015. Ecological data sharing. Ecol. Inf. 29, 33–44.
Midway, S.R., Wagner, T., Zydlewski, J.D., Irwin, B.J., Paukert, C.P., 2016. Transboundary fisheries science: meeting the challenges of inland fisheries management in the 21st century. Fisheries 41 (9), 536–546.
Midway, S., Robertson, M., Flinn, S., Kaller, M., 2020. Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. PeerJ 8, e10387.
NCBI Resource Coordinators, 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44 (D1), D7–19. https://doi.org/10.1093/nar/gkv1290. Jan 4. (Epub 2015 Nov 28. PMID: 26615191; PMCID: PMC4702911).

Nichols, J.D., Williams, B.K., 2006. Monitoring for conservation. Trends Ecol. Evol. 21 (12), 668–673.
Peters, D.P., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5 (6), 1–15.
Reichmann, O.J., Jones, M.B., Schildhauer, M.P., 2011. Challenges and opportunities of open data in ecology. Science 331 (6018), 703–705.
Soranno, P.A., Cheruvelil, K.S., Bissell, E.G., Bremigan, M.T., Downing, J.A., Fergus, C.E., Filstrup, C.T., Henry, E.N., Lottig, N.R., Stanley, E.H., Stow, C.A., 2014. Cross-scale interactions: quantifying multi-scaled cause–effect relationships in macrosystems. Front. Ecol. Environ. 12 (1), 65–73.
Soranno, P.A., Bissell, E.G., Cheruvelil, K.S., Christel, S.T., Collins, S.M., Fergus, C.E., Filstrup, C.T., Lapierre, J.F., Lottig, N.R., Oliver, S.K., Scott, C.E., 2015a. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. GigaScience 4 (1) (s13742–015).
Soranno, P.A., Cheruvelil, K.S., Elliott, K.C., Montgomery, G.M., 2015b. It's good to share: why environmental scientists' ethics are out of date. BioScience 65 (1), 69–73.
Whittier, J., Sievert, N., Loftus, A., Defilippi, J.M., Krogman, R.M., Ojala, J., Litts, T., Kopaska, J., Eiden, N., 2016. Leveraging BIG data from BIG databases to answer BIG questions. Fisheries 41 (7), 417–419.