A lack of open data standards for large infrastructure projects hampers social-ecological research in the Brazilian Amazon (#110849)

First submission

Guidance from your Editor

Please submit by **27 Jan 2025** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the materials page.

9 Figure file(s)
1 Table file(s)

Custom checks

Human participant/human tissue checks

- Have you checked the authors ethical approval statement?
- Does the study meet our article requirements?
- Has identifiable info been removed from all files?
- Were the experiments necessary and ethical?

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- **Impact and novelty is not assessed.** Meaningful replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

A lack of open data standards for large infrastructure projects hampers social-ecological research in the Brazilian Amazon

Jacy L Hyde [Equal first author, 1], **Christine Swanson** [Corresp., Equal first author, 1, 2], **Stephanie A Bohlman** [1, 3], **Simone Athayde** [4], **Emilio M Bruna** [5, 6], **Denis R Valle** [1]

[1] School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, Florida, United States

[2] Division of Undergraduate Education, U.S. National Science Foundation, Alexandria, Virginia, United States

[3] Smithsonian Tropical Research Institute, Balboa, Ancón, Panama

[4] World Resources Institute, Washington, District of Columbia, United States

[5] Center for Latin American Studies, University of Florida, Gainesville, Florida, United States

[6] Department of Wildlife Ecology & Conservation, University of Florida, Gainesville, Florida, United States

Corresponding Author: Christine Swanson
Email address: acswanso@nsf.gov

New infrastructure projects are planned or under construction in several countries, including in the bioculturally diverse Amazon, Mekong, and Congo regions. While infrastructure development can improve human health and living standards, it may also lead to environmental degradation and social change. Accessible, high quality data about infrastructure projects is essential for both monitoring these projects and studying their social and environmental impacts. We investigated the availability and quality of data on infrastructure projects in the Brazilian Amazon by reviewing the academic literature and surveying researchers from the conservation and development community. We used the results of these surveys to identify recommended steps for the gathering, organizing, and sharing of infrastructure data by social-ecological researchers and practitioners.

Although data on infrastructure in the Brazilian Amazon were generally available, they were often of poor quality and lacked information critical for monitoring and research. Data were often difficult to find and reformat, resulting in loss of time and resources for researchers and other stakeholders. Discrepancies between researchers' survey responses on data needs and the types of data used in peer-reviewed articles on infrastructure projects indicate the following information was often missing: geographic extent of the project, construction and operation dates, and project type (e.g., paved vs unpaved road). Including these data in a standardized format, along with making them more readily accessible by hosting them in public repositories and ensuring they are current and comprehensive, would facilitate research and improve planning, decision-making, and monitoring of existing and future infrastructure projects in Brazil and other developing countries.

A lack of open data standards for large infrastructure projects hampers social-ecological research in the Brazilian Amazon

Hyde, J.L[a]., Swanson, A.C. [a,b*], Bohlman, S.A. [a,c], Athayde, S.[d], Bruna, E.M.[e,f], Valle, D.R. [a]

a. *School of Forest, Fisheries, and Geomatics Sciences, University of Florida, 136 Newins-Zeigler Hall, Gainesville, FL, 32611 USA.*
b. *Division of Undergraduate Education, Directorate for STEM Education, U.S. National Science Foundation. 2415 Eisenhower Avenue, Alexandria, VA 22314, USA.*
c. *Smithsonian Tropical Research Institute, Balboa, Ancón, Panama*
d. *World Resources Institute (WRI), 10 G Street NE, Washington, DC 20002, USA.*
e. *Center for Latin American Studies, University of Florida, 319 Grinter Hall, P.O. Box 115530, Gainesville, FL, 32611-5530, USA.*
f. *Department of Wildlife Ecology & Conservation, University of Florida, 110 Newins-Zeigler Hall, Gainesville, FL, 32611 USA*

Corresponding Author:
A. Christine Swanson
acswanso@nsf.gov
2415 Eisenhower Avenue
Alexandria, VA 22314

Highlights

- Infrastructure projects across the globe spur economic development but also lead to social-ecological degradation.
- Researchers and practitioners need current and comprehensive data to better understand and mitigate social-ecological changes from infrastructure.
- Infrastructure data are often unavailable, inaccessible, or incomplete.
- Finding and organizing datasets cost researchers hundreds of hours and may lead to abandoned projects.
- To promote better research outcomes, governments and NGOs should ensure datasets on infrastructure are accessible, current, comprehensive, and include such vital information as the project's geographic extent, dates of construction and operation, project type, and essential technical data.

Abstract

New infrastructure projects are planned or under construction in several countries, including in the bioculturally diverse Amazon, Mekong, and Congo regions. While infrastructure development can improve human health and living standards, it may also lead to environmental degradation and social change. Accessible, high quality data about infrastructure projects is essential for both monitoring these projects and studying their social and environmental impacts. We investigated the availability and quality of data on infrastructure projects in the Brazilian Amazon by reviewing the academic literature and surveying researchers from the conservation and development community. We used the results of these surveys to identify recommended steps for the gathering, organizing, and sharing of infrastructure data by social-ecological researchers and practitioners.

Although data on infrastructure in the Brazilian Amazon were generally available, they were often of poor quality and lacked information critical for monitoring and research. Data were often difficult to find and reformat, resulting in loss of time and resources for researchers and other stakeholders. Discrepancies between researchers' survey responses on data needs and the types of data used in peer-reviewed articles on infrastructure projects indicate the following information was often missing: geographic extent of the project, construction and operation dates, and project type (e.g., paved vs unpaved road). Including these data in a standardized format, along with making them more readily accessible by hosting them in public repositories and ensuring they are current and comprehensive, would facilitate research and improve planning, decision-making, and monitoring of existing and future infrastructure projects in Brazil and other developing countries.

Keywords: open data, infrastructure, social-ecological research, conservation, tropics, Brazil, Amazon

1. Introduction

Access to comprehensive, high-quality infrastructure project data is critical to studying, monitoring, and mitigating the social-ecological impacts of infrastructure (Joppa et al., 2016). This is particularly important given the millions of roads, dams, hydroways, ports, transmission lines and other major infrastructure projects that are currently operational, under construction, or planned worldwide, including planned massive regional, national, or multi-national infrastructure expansions (e.g., the Initiative for the Integration of the Regional Infrastructure of South America (IIRSA)[1], and China's Belt and Road Initiative[2]). Governments, nongovernmental organizations, and project funders collect and make available such data to monitor compliance and assess environmental impacts (Ciborra, 2005). Transparency and accountability resulting from making data available throughout the development process may help minimize inefficiency, corruption, and the mismanagement of public construction projects that have resulted in an annual loss of $4 trillion globally (Transparency and Accountability Initiative, 2014). Researchers can help improve estimates of trade-offs and impacts for various project alternatives (Laurance et al., 2015), and third parties may bring innovative ideas and solutions to the table (Janssen et al., 2012). They can also hold the government accountable for including social-environmental variables in licensing or construction decisions and developing adequate consultation and compensation processes for affected populations (Moran et al., 2018; Pereira, 2021; Transparency and Accountability Initiative, 2014). While the expansion and improvement of infrastructure can help increase standards of living and improve human health (Brenneman & Kerf, 2002; Calderón & Servén, 2004; Estache, 2003; Johansson & Goldemberg, 2002; Martínez & Ebenhack, 2008; Slough et al., 2015), large infrastructure projects can also lead to environmental degradation (Laurance, 2018; Laurance & Arrea, 2017; Pfaff et al., 2018; C. M. Souza et al., 2019) and negatively impact Indigenous peoples and other local communities (Arrifano et al., 2018; Fearnside, 1999; Gauthier & Moran, 2018). These social-ecological impacts cannot be properly identified and quantified without information about the infrastructure projects themselves.

[1] http://www.iirsa.org/

[2] http://english.gov.cn/beltAndRoad/

Providing access to data and information about public or public-private infrastructure projects is typically the responsibility of a government institution. Government agencies are usually responsible for planning projects for licensing or other administrative purposes and for monitoring existing projects and, thus, should have relevant information about these projects. Many countries have access-to-information laws requiring the release of information to the public (Kaufmann & Bellver, 2005; Relly, 2010) or have signed transparency pledges. Full implementation of these policies is rare, however, due in part to resource and technological constraints, lack of motivation and capacity among agencies, or unclear designation of responsibility (Attard et al., 2015; Ciborra, 2005; Di Ciommo, 2015; Janssen et al., 2012; Wang & Lo, 2016). Consequently, data about public or public-private infrastructure projects are often unavailable (Attard et al., 2015).

Even when infrastructure data are available, they may not be of a sufficient quality for specific social-ecological research. Task-independent data quality standards, which have been proposed by several entities, apply to datasets independent of research question or usage and focus on the completeness, accuracy, and currency of information (Open Data Charter, 2015; Pipino et al., 2002; Vetrò et al., 2016). These standards do not provide guidance on what information should be included within a dataset. However, for nearly all research projects, data quality is defined as the degree of usefulness in a particular task or context and is highly dependent on the user (Stvilia et al., 2007), requiring context-specific content. Consequently, even when data do meet task-independent quality standards, the dataset may still be of limited use because it lacks correct or sufficient information to guide a specific decision or to enable a specific task. For example, basic information about project location or date of construction is not always readily available to researchers, requiring them to invest significant time and resources searching for or collecting these data (Hyde et al., 2018; Klarenberg et al., 2019; Tucker Lima et al., 2016)

Accessibility to high quality data is especially important in countries undergoing rapid infrastructure development. Brazil is a culturally and ecologically hyperdiverse country, but major infrastructure development plans throughout the country, and especially in the Amazon, threaten this diversity (Athayde et al., 2019). Brazil has a strong legal framework to promote transparency, codifying the right to access information in the 1988 Brazilian Constitution and reinforcing this right with various national and international laws, ordinances, and supporting institutions[3]. In 2011, Brazil cofounded the Open Government Partnership (OGP)[4], which seeks to promote transparency, empower citizens, fight corruption, and harness new technologies to strengthen governance. Unfortunately, most Brazilian government portals are not in compliance with international open government data criteria (Di Ciommo, 2015). It is also unknown whether data standards are followed throughout all sectors or if the standards are adequate for conducting meaningful research into social-ecological impacts of development.

Guidelines for the content of infrastructure datasets may improve the usefulness of these datasets for social-ecological research by ensuring they contain certain critical attributes and information (Joppa et al., 2016). Using infrastructure development in the Brazilian Amazon as a case-study, we conducted a systematic review of how infrastructure data have been used in social-ecological research in academic publications. Specifically, we asked: 1) What data are

[3] https://www.right2info.org/recent/access-to-public-information-in-brazil-what-will-change-with-law-no.-12.527-2011

[4] www.opengovpartnership.org

required for social-ecological research related to infrastructure projects? 2) How accessible and complete are public datasets on infrastructure projects in the Brazilian Amazon? We then surveyed practitioners and researchers about their data needs and efforts in searching for and using data on infrastructure. We used the results of our literature review and survey to identify what attributes should be included in all infrastructure datasets to maximize the utility of these datasets for researchers and other interested parties. Finally, we evaluated two datasets available on Brazilian government websites to determine how well they conformed to task-independent standards and whether they included the critical attributes we identified. While our study focuses on open data from Brazil, our recommendations are broadly applicable to infrastructure and development projects across the world.

2. Materials & Methods

2.1 Systematic literature review

We performed a systematic literature review to determine what information researchers have used when assessing the social and ecological impacts of infrastructure projects in the Brazilian Amazon. We only considered studies published after the passage of the Brazilian Federal Access to Information Law in 2011 (Lei. 12.527/201[5]). In accordance with this law, after 2011, scientists should have been able to acquire government data on infrastructure projects from Brazilian government agencies if this policy was fully enacted. We performed the literature review using the Web of Science's[6] "Core Collection" in September and October 2018. The search strings we used for the literature review are available at https://doi.org/10.5281/zenodo.10626908. We specifically looked for studies focusing on environmental impacts, management, or conservation in relation to current or planned infrastructure projects in the Brazilian Amazon. We only included studies if they specifically used some type of infrastructure dataset or information in their analysis or required infrastructure data to plan the research study. For each study, we determined the type of information used about infrastructure projects and focused on the project attributes (e.g., construction date, location, budget, etc.). We recoded the citation, topic, academic discipline, infrastructure type, the dataset(s) and types of data used, and the infrastructure attributes for each study.

2.2 Key informant survey

To understand data needs and experiences, we surveyed key researchers and practitioners who focus on social and/or ecological topics from a list of the corresponding authors of the

[5] Brasil. Lei ordinária nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal. Diário Oficial da União 2011; 18 nov.

[6] The Web of Science (WOS), previously known as Web of Knowledge, is an online subscription-based scientific citation indexing service that provides a comprehensive citation search. The Web of Science Core Collection consists of six online databases: Science Citation Index; Social Sciences Citation Index. Arts and Humanities Citation Index; Emerging Sources Citation Index; Book Citation Index; and Conference Proceedings Citation Index. Additional databases available in WOS searches include SciELO Citation Index; BIOSIS Citation Index; MEDLINE1; CABI; and Zoological records. Website: https:// clarivate.com/products/web-of-science/ Source: Wikipedia: https://en. wikipedia.org/wiki/Web_of_Science,

papers in the literature review, members of the Amazon Dams International Research Network[7] (ADN; Athayde et al., 2019), and members of the Governance and Infrastructure in the Amazon[8] (GIA) working group (Mere-Roncal et al., 2021). The ADN and GIA coordinate social-ecological research and information-sharing about infrastructure in the Amazon and are comprised of researchers, NGO practitioners, and members of government agencies. With the survey, we collected demographic information and asked participants questions about types of infrastructure projects for which they searched, the information required about these projects, where they searched for information, how long it took to find relevant information and format it for use, what they did if they could not find appropriate data, and about data quality based on task-independent standards (Vetrò et al., 2016). Finally, we asked participants to list and rank infrastructure project attributes that were important for their use. This survey was approved by the University of Florida's Institutional Review Boards (IRB #B201600928). Respondents gave written consent to participate in the survey. The full survey is available at https://doi.org/10.5281/zenodo.10626908.

From the survey responses, we summarized which attributes were most important across infrastructure projects. We compared the data survey respondents wanted to data used in the literature and considered discrepancies between the two sources a possible data gap where necessary data might not be available. We evaluated the data quality and the amount of effort spent on data gathering, cleaning, and formatting by performing summary statistics on survey responses. To examine differences in data quality between data retrieved from government versus non-government sources, we only considered answers from participants who reported retrieving data exclusively from a government repository or exclusively from a non-government repository. We also combined responses for all non-government sources (i.e., academic, NGO, other).

2.3 Proposing critical attributes for infrastructure data sets

Based on attributes used in the literature review and survey participants' rankings of attribute importance, we created context-specific standards for infrastructure datasets that are complimentary to the task-independent data quality standards. We considered attributes that were ranked in the top five in the key informant survey more than 40% of the time as critical for inclusion in infrastructure datasets. By identifying these critical attributes, we strived to encourage the availability of information required to conduct social-ecological research about infrastructure projects.

2.4 Evaluating available infrastructure datasets

To further understand the quality of open data on infrastructure from the Amazon region, we evaluated two infrastructure datasets on whether they contained the attributes we identified as critical for social-ecological research and whether they complied with the task-independent framework provided by Vetrò et al. (2016). Our test cases were large dams and roads as they are drivers of social-ecological change in the Amazon (Chen et al., 2015; Latrubesse et al., 2017; Laurance & Arrea, 2017; Nepstad et al., 2001) and frequently appeared in our survey responses and literature review. Therefore, it is especially important that these data are of high quality and useable for social-ecological research.

[7] http://amazondamsnetwork.org

[8] https://giamazon.org

We downloaded data on May 22, 2019 from the agencies that oversee the dams and roads, the Agencia Nacional de Energia Elétrica (ANEEL)[9] and the Departamento Nacional de Infraestrutura de Transportes (DNIT)[10], respectively. We evaluated the quality of these two publicly available infrastructure datasets based on inclusion of information we identified as critical (see previous paragraph) and five characteristics from Vetrò et al.'s (2016) task-independent framework: (1) accuracy of spatial components, (2) completeness, (3) currency (up-to-date), (4) machine-readability, and (5) metadata quality. We assessed the spatial accuracy of the datasets by randomly selecting 50 existing projects in each dataset and verifying their locations in Google Earth using the same map projection. If the project was within 30 m (the size of a Landsat pixel) of the location listed on the dataset, it was considered spatially accurate. We quantified how current the dataset was based on the date of the last update. Completeness was difficult to assess because it was unclear in many of the columns whether an empty cell was purposefully empty (the metadata did not provide this information). Instead of scoring the whole data set based on completeness, we chose the first two columns in each data set that were understandable without metadata and that clearly should have been complete, and we determined the percent of empty cells in these columns. Metadata quality was used as proxy for traceability (which measures the history of the data set) and understandability, both of which are somewhat subjective. Thus, we considered the metadata complete if it was present and contained explanations of the attributes in the data, its author, the geographic coordinate system of the shapefile, and the publication date.

3. Results

3.1 Systematic literature review

Sixty-two studies fit our criteria for the systematic literature review of articles that have investigated social-ecological impacts of infrastructure in the Brazilian Amazon. Together, the articles used infrastructure data 94 times, requiring 236 attributes about those infrastructure projects (Figure A-1B). Hydropower projects were the most common infrastructure category investigated (43 datasets), followed by roads and highways (18 datasets). By far the most used attribute about infrastructure projects was the geographic location of the project (66 times). The project name, its full geographic extent, and basic technical information were also used frequently (Figure A-1C). The articles focused on a range of topics, most frequently on social issues (such as displacement, livelihoods, socio-environmental conflict, human health, etc.), land use/land cover change, and aquatic ecology (Figure A-1A).

3.2 Key informants survey

From the 472 people we contacted, a total of 87 people responded to the survey, with 68 completions (response rate = 18.4%, completion rate = 14.4%). Most participants (61.8%) were located in Brazil and 70.6% of respondents were in some stage of an academic career (Figure A-2). Participants were primarily researching socioeconomic topics (20.6%), land use/land cover change (20.0%), Indigenous peoples (16.3%), and natural resources (13.2%) (Figure A-2).

There were 539 instances for which data on infrastructure were used by the survey participants. Combined, hydroelectric dams (28.8%), small dams (14.6%), and roads and highways (9.6%) accounted for more than half of the data searches (Figure A-2). Participants

[9] https://sigel.aneel.gov.br/Down/

[10] https://www.dnit.gov.br/planejamento-e-pesquisa/dnit-geo

searched for information relatively evenly across project phases: 37.5% searched for information about the planning phase, 34.1% the construction phase, and 28.4% the post-construction phase. Although participants searched for a wide range of information about the infrastructure projects; point location (10.5%), name (8.7%), status (8.4%), construction and operation dates (8.3% each), and full geographic extent (8.3%) were the most sought-after data attributes (Figure A-3).

3.3 Comparison of attributes in literature review versus survey

There were substantial gaps between the data attributes ranked within the top five needed for research by survey participants compared to the frequency of use these attributes in articles we reviewed (Figure 1). Most attributes were ranked within the top five attributes required for social-ecological analysis in the survey at a higher rate than they were used in the studies we reviewed, including the geographic extent of the project, construction and operation dates, and project name and status. In contrast, point location was used more often in the literature than it was ranked in the top five attributes, possibly indicating this attribute was more available than the geographic extent of the infrastructure project, which may have been a more useful attribute. Combined, these results highlight potential gaps in infrastructure data availability.

3.2.2 Data accessibility and quality

Government sources were the most common place to search for information (39%), but academic and NGO sources were also frequently queried (31% and 24%, respectively) (Figure 2). Of the 188 searches for government data, 82% datasets were found from government sources, 14% required additional searches on non-government sources to access the data, and only 3.7% were not found at all. Respondents reported successful access of data from academic and NGO sources in 66% and 64% of the attempts, respectively, a lower rate than from government sources. For academic and NGO sources, 2% were not found (Figure 2).

Survey participants reported high uncertainty about accuracy of non-spatial components of data. Thirty-nine percent of respondents reported either that non-spatial data had low accuracy or that the respondent could not evaluate the accuracy (Figure 3A). Conversely, all participants reported that spatial accuracy was at least moderate in quality (Figure 3A). Accuracy ratings were similar for data obtained from both government and non-government sources (Figure A-4). Respondents rated data sets low for task-independent standards (Figure 3B). The highest scoring category of task-independent standards was machine readability, although one-third of the data sets scored low in in this category. Other categories scored even worse, with 45%, 65% and 88% of the respondents rating currentness, metadata quality and completes, respectively, as low (Figure 3B). Data acquired from government data sources scored higher in terms of task-independent data quality compared to non-government sources (Figure A-4). For example, almost 75% of the government-sourced data was machine readable, compared to only 40% of non-government data (Figure A-4).

Most respondents (over 95%) were able to find data (Figure 4). Ten respondents were unable to find the data they required (Figure 4) for a variety of infrastructure types: large hydroelectric dams (1), railroads (1), roads (1), solar energy plants (1), transmission/distribution lines (1), waterways (1), wastewater/sewage (2), and small dams (2). Five of the respondents who were unable to find data abandoned their projects altogether, two used proxy datasets, two collected the data themselves, and a third respondent unsuccessfully attempted to collect data.

The time spent searching for data showed a bimodal distribution. For data sets that were found, 35% of respondents reported spending less than eight hours searching for data before

finding it, while almost 30% spent more than 168 hours in their search for data (Figure 4). This is the equivalent of more than one month's worth of work, assuming a 40-hour work week.

Even when respondents were able to find data, 90% had to spend additional time putting those data into a usable format (Figure 4). Time spent formatting data also showed a bimodal distribution. Thirty-five percent of respondents spent less than eight hours formatting the data whereas, 37% spent at least 168 hours to make data useable (Figure 4). Twenty respondents reported never being able to get their data into a usable format. The unusable datasets varied in quality. Three were incomplete, eight were not current, four were not machine readable, and six had low quality metadata. Time spent searching and formatting data was similar between government and non-government sources (Fig. A-5).

3.3 Critical attributes for infrastructure datasets

Based on the feedback from the literature review and surveys, we propose a list of critical attributes that should be included in all infrastructure datasets. At a minimum, all infrastructure datasets should include:

1. The project name
2. Spatial extent of the project
3. Basic technical information about the project, which would vary by project but may include capacity (electrical or physical), voltage, bandwidth, number of beds, number of students, etc.
4. Date construction started on the project
5. Project type, which also varies by infrastructure type but may also include paved vs. dirt road, run-of-river vs. impoundment dam, primary vs. secondary school, highway vs. access road, etc.
6. Date project began operations

3.4 Evaluation of public datasets

We evaluated whether two publicly available infrastructure datasets (large hydroelectric dams and roads) contained the critical attributes described above and assessed the quality of these datasets based on the task-independent standards. Though we were not able to find either dataset on Brazil's central data repository (dados.gov.br), the datasets were available from the websites of the agencies that oversee dams (Agencia Nacional de Energia Elétrica) and roads (Departamento Nacional de Infraestrutura de Transportes). Neither dataset contained all the critical attributes we proposed (Table 1). While both datasets did include project names, neither included project construction or operation dates. The dams dataset included point data and technical information on reservoir sizes but did not include the geographic extent of reservoirs or the dam buildings. The roads dataset included project type (e.g., paved/unpaved, federal/state) and geographic extent in the form of line features for the full length of roads but lacked technical information.

The datasets had high spatial accuracy with 96% (dams) and 98% (roads) of randomly-selected points falling within 30 m of their location based on satellite imagery. To measure completeness, we determined the number of filled cells within the first two easily understandable columns: project name and owner for dams, and project name and status for roads. For the dams dataset, 94.5% of cells within these two columns were filled while 100% of the cells within the two roads columns were filled. Across all columns, 17.5% (dams) and 21.7% (roads) of cells

were missing data, but it was unclear whether these cells were supposed to be empty because there were no attribute descriptions in the metadata. The dams dataset was current, but there was no information on the currency of the roads dataset (Table 1). Both datasets were available in machine readable formats (ESRI shapefile or KMZ). The metadata quality was low for both datasets. The roads dataset included no metadata for author, date of creation or latest update, attribute description, or geographic datum. The metadata for the dams dataset only included information about the date of latest update (Table 1).

4. Discussion

4.2 Data quality and availability for the Amazon

While we found that infrastructure data for projects in the Legal Amazon were generally available, finding these data often required extensive, time-consuming searches. Furthermore, data sets were often low quality based on task-independent standards, which required many users to spend additional time on data formatting. Studies of publicly available data of non-infrastructure data have shown similar issues in accessibility and usability (Roche et al., 2015; Vines et al., 2014) indicating a broad need for greater data accessibility across disciplines. The federal government is often the main regulator, if not the main funder and co-owner, of large-scale infrastructural projects, and government repositories were the most common place where stakeholders searched for these data. Therefore, it is especially important that government agencies provide easily accessible and high-quality data for projects under their jurisdiction. Unfortunately, our survey reveals that not all government data is easily accessible or interpretable. For example, much of the time participants spent searching for data may have been spent looking across multiple government websites and jurisdictions for desired data and/or parsing difficult-to-navigate websites (de Oliveira & Silveira, 2018a). A central repository to host infrastructure data, either provided by the federal government or by non-government institutions, would likely reduce the search time thereby increasing the accessibility of information. Examples of non-governmental central repositories that store infrastructure and environmental data include Global Forest Watch and MapBiomas. These and other central repositories could serve to collate data across jurisdictions.

Our results demonstrate some of the costs of poor accessibility and quality. After previously investing time into searching for data (and in some cases, putting them into usable formats), poor quality or missing data resulted in 12 abandoned projects. Alternatively, some researchers invested time and resources to collect new data, thereby having less time and resources for other projects. The cost of poor quality or inaccessible data in time and financial resources are more easily inferred from our survey, but it is challenging to determine the costs associated with failing to generate potentially crucial information for assessing and managing impacts and planning future projects more sustainably. Delays or failures to create this information could have long-term consequences during the global infrastructure boom.

4.2 Gaps in data availability

The discrepancy between the desired attributes (from the survey) and previously used attributes (from the literature review) may indicate an important data accessibility gap, with project type, construction and operation dates, and project status being the least accessible, as represented in published studies, but highly demanded attributes in the surveys. The lack of and/or poor quality of these attributes was confirmed by our analysis of the publicly available roads and dams datasets. Similarly, point location was used far more often than the geographic

boundaries of the project in published studies, although both attributes were requested equally in the survey, likely indicating that the full geographic extent of projects is less available. The failure to include critical desired infrastructure information likely impacts the details of research and monitoring being conducted. For example, utilizing dates of construction, operation and changes in project status may allow more nuanced analysis of the timeline of impacts in contrast to using a single date to determine pre- and post-project impacts. Similarly, using geographic extent rather than a point location for an infrastructure project may be important to accurately determine the spatial extent of direct and indirect impacts of that project, such as deforestation, land use change, and displacement.

4.3 Accuracy of assessed data

Our review of the public data sets for dams and roads from government websites revealed that while Brazil appears to follow the laws regarding access to information by freely providing basic information about infrastructure projects, the overall task-independent and conservation-specific quality could be substantially improved to increase the usefulness of these data. Neither dataset contained all critical data we identified in our study. Both had low-quality metadata, which made it challenging to interpret many of the attributes. Furthermore, external verification of the data sets was difficult. For example, it was impossible to verify accuracy of the non-spatial data about the infrastructure projects without exhaustively searching planning documents, a task which would have taken dozens to hundreds of hours to complete. This uncertainty is also reflected in the frequent "I don't know" responses from our survey participants about the accuracy of non-spatial components. Validation of spatial accuracy may be more easily accomplished through use of satellite imagery or other remote sensing data, but this validation is still a time-consuming endeavor. As an example, over six months (Hyde et al., 2018) hand-digitized all the transmission lines in the Legal Amazon from satellite data and compared them to two public datasets, which differed from each other. Only one was spatially accurate and neither dataset included every transmission line in the region. This demonstrates the importance of external validation of spatial and non-spatial accuracy to ensure data quality. One way to increase confidence in public datasets and reduce time spent validating data before use would be to develop a system allowing users to rate completeness and accuracy of these datasets.

4.4 Critical data attributes for social-ecological research

Previous studies illustrate how the attributes we identified as being critical for understanding impacts of infrastructure have been used. For example, Swanson & Bohlman (2021) used the names, operations dates, dam types (run-of-river vs. impoundment), and geographic extent of reservoirs to quantify changes in land cover in the Tocantins River watershed after the installation of multiple hydropower dams. Nickerson et al. (2022) compared deforestation surrounding large hydropower dams and small dam clusters in the Legal Amazon, which required information about the construction and operation dates as well as technical information and type of dams. From an energy planning perspective, de Faria & Jaramillo (2017) investigated alternatives to hydropower expansion in the Amazon using location and capacity data on all current and planned wind, solar, thermal, and hydropower in the region. Additionally, de Souza et al. (2023) used data for highways, waterways, railways, and ports, including project type, geographic extent, and technical information, to model port hinterlands in the Brazilian Amazon. Finally, Menezes et al. (2018) modeled vulnerability of Amazonian municipalities to climate change using technical information about hospitals. These studies demonstrate the

importance of information about specific infrastructure projects to monitor social-ecological change related to new development and to improve planning and monitoring efforts for future infrastructure development. They also illustrate that the content-specific attributes recommended for inclusion in infrastructure datasets are ubiquitous across infrastructure types.

Though this study focused on the Brazilian Amazon, the information needed from infrastructure data is likely to the same in other geographic regions. For instance, Lupinetti-Cunha et al. (2022) and Tisle[illegible] al. (2022) both used the geographic extent and project type to model the effects of roadless areas on land cover change and conservation across all of Brazil, not just the Amazon. Beyond Brazil, Flecker et al. (2022) used location and technical information (capacity in MW) about hydropower dams to investigate how environmental impacts from damming could be reduced across the Amazon basin. Baird et al. (2021) used information including dam names, operation and construction dates, spatial extent, and technical information to investigate downstream impacts of dams and the need for Indigenous and traditional knowledge to mitigate those impacts in the Amazon, Canada, Laos, and Vietnam. Finally, Ding et al. (2022) used information about project type to model the carbon emissions of 5G cell stations in China. These studies illustrate that the information identified in our literature review and through our key informant survey is applicable beyond the geographic boundaries of the Brazilian Legal Amazon. Though we developed these recommendations based on research conducted in the Brazilian Amazon, our recommendations were designed to be general enough to be applicable to infrastructure development initiatives across the globe and to improve the usability of infrastructure data for a broad range of research initiatives.

4.5 Caveats

While considerable effort was made to obtain a representative sample of stakeholders interested in all infrastructure types in the Amazon region, the survey population was biased toward a hydropower focus. However, we note that the literature review results also were also skewed towards dams. Thus, this may simply reflect the general focus of the scientific community on dams in the Amazon region in the context of the hydropower boom (Athayde et al., 2019). Despite the focus on dams, at least one person searched for every attribute for every infrastructure type, so we believe that our results and the critical data attributes can be used broadly across infrastructure types. Finally, most of our survey respondents held academic positions, so these results may reflect the needs of the academic research community more strongly than those of government, the private sector, or NGO communities.

4.6 Conclusion and future directions

Access to high quality data about infrastructure projects has the potential to improve the quality and efficiency of social-ecological research and assessment of impacts related to new and planned development projects. With access to the best data available, third parties and governments can ensure the accuracy and accountability of environmental impact assessments (Laurance et al., 2015), fair compensation to impacted communities, and appropriate mitigation plans (Hunter et al., 2021). They can also inform improved watershed-level planning, as opposed to the project-by-project planning (Athayde et al., 2019), as well as better identify projects that are more harmful than helpful. Data transparency may also help reduce corruption in the infrastructure sector (Kaufmann & Bellver, 2005; Ruijer et al., 2017).

To achieve these goals, governments and third parties would need to release datasets that conform to task-independent standards and contain, at minimum, the critical attributes identified

in this study. In cases where governments have not or cannot provide accessible and high-quality data, the research and NGO communities can be important sources of data to fill in these gaps. In addition, it is important that individual researchers share data on infrastructure that they collect on public archives. A culture of open data will reduce the redundant collection of data while allowing researchers to be credited for their work through citations (Allen & Mehler, 2019). The research and government communities should also strive to remove barriers to accessibility by investing in comprehensive and up-to-date central repositories to host this data. As countries continue to expand and update their infrastructure, promoting transparency and data sharing about the projects is an important step in implementing the right to access to information, as well as improved public participation in decision-making related to current and future infrastructure.

Acknowledgments

We would like to thank the University of Florida (UF) Water Institute, the School of Forest, Fisheries and Geomatics Sciences and the UF Tropical Conservation and Development Program (TCD) for their support, as well as Dr. Stephen Perz, Dr. David Kaplan, and the 2015 Water Institute Graduate Fellows. Dr. Charles Jekel was instrumental in helping with coding issues. We are grateful to our colleagues in the Amazon Dams International Research Network/Rede Internacional de Pesquisa em Barragens Amazônicas/ Red Internacional de Investigación en Represas Amazónicas (ADN/RBA/ RIRA), to the participants of the civil society initiative and working group GT Infraestrutura, and to our associates at the Agência Nacional de Energia Elétrica (ANEEL) and Empresa de Pesquisa Energética (EPE) for their input and support during this research process.

Competing Interests

The authors have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

J.L. Hyde: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review & editing, project administration; **A.C. Swanson:** software, validation, formal analysis, data curation, writing – review & editing, visualization, project administration; **S. Bohlman:** resources, writing – review & editing, supervision, funding acquisition; **S. Athayde:** writing – review & editing, supervision, funding acquisition; **E.M. Bruna:** writing – review & editing, supervision; **D.R. Valle:** conceptualization, methodology, writing - review & editing, supervision, funding acquisition

Funding: This research is partially based upon work supported by the U.S. National Science Foundation (NSF) under Grant No. 1617413. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect NSF views. Support was also provided by USDA National Institute of Food and Agriculture, McIntire Stennis project 1024612 to S. A. Bohlman. The survey was performed under the IRB #B201600928. J.L. Hyde and A.C. Swanson were supported by the University of Florida Water Institute. A.C. Swanson had additional funding from the University of Florida Informatics Institute, NASA FINESST award #80NSSC19K1355, and the NSF. Part of this research was performed while A.C. Swanson held a National Research Council Research Award at the U.S. Naval Research Laboratory.

526 **References**

527 Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career
528 and beyond. *PLOS Biology*, *17*(5), e3000246.
529 https://doi.org/10.1371/journal.pbio.3000246
530 Arrifano, G. P. F., Martín-Doimeadios, R. C. R., Jiménez-Moreno, M., Ramírez-Mateos, V., da
531 Silva, N. F. S., Souza-Monteiro, J. R., Augusto-Oliveira, M., Paraense, R. S. O., Macchi,
532 B. M., do Nascimento, J. L. M., & Crespo-Lopez, M. E. (2018). Large-scale projects in
533 the amazon and human exposure to mercury: The case-study of the Tucuruí Dam.
534 *Ecotoxicology and Environmental Safety*, *147*, 299–305.
535 https://doi.org/10.1016/j.ecoenv.2017.08.048
536 Athayde, S., Mathews, M., Bohlman, S., Brasil, W., Doria, C. R., Dutka-Gianelli, J., Fearnside,
537 P. M., Loiselle, B., Marques, E. E., Melis, T. S., Millikan, B., Moretto, E. M., Oliver-
538 Smith, A., Rossete, A., Vacca, R., & Kaplan, D. (2019). Mapping research on
539 hydropower and sustainability in the Brazilian Amazon: Advances, gaps in knowledge
540 and future directions. *Current Opinion in Environmental Sustainability*, *37*, 50–69.
541 https://doi.org/10.1016/j.cosust.2019.06.004
542 Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government
543 data initiatives. *Government Information Quarterly*, *32*(4), 399–418.
544 https://doi.org/10.1016/j.giq.2015.07.006
545 Baird, I. G., Silvano, R. A. M., Parlee, B., Poesch, M., Maclean, B., Napoleon, A., Lepine, M., &
546 Hallwass, G. (2021). The Downstream Impacts of Hydropower Dams and Indigenous and
547 Local Knowledge: Examples from the Peace–Athabasca, Mekong, and Amazon.
548 *Environmental Management*, *67*(4), 682–696. https://doi.org/10.1007/s00267-020-01418-
549 x

Brenneman, A., & Kerf, M. (2002). *Infrastructure & Poverty Linkages*. World Bank. https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@ed_emp/@emp_policy/@invest/documents/publication/wcms_asist_8281.pdf

Calderón, C., & Servén, L. (2004). *The effects of infrastructure development on growth and income distribution* (No. 3400; Policy Research Working Paper). World Bank. https://doi.org/10.1596/1813-9450-3400

Chen, G., Powers, R. P., de Carvalho, L. M. T., & Mora, B. (2015). Spatiotemporal patterns of tropical deforestation and forest degradation in response to the operation of the Tucuruí hydroelectric dam in the Amazon basin. *Applied Geography*, *63*, 1–8. https://doi.org/10.1016/j.apgeog.2015.06.001

Ciborra, C. (2005). Interpreting e-government and development: Efficiency, transparency or governance at a distance? *Information Technology and People*, *18*(3), 260–279. https://doi.org/10.1108/09593840510615879

de Faria, F. A. M., & Jaramillo, P. (2017). The future of power generation in Brazil: An analysis of alternatives to Amazonian hydropower development. *Energy for Sustainable Development*, *41*, 24–35. https://doi.org/10.1016/j.esd.2017.08.001

de Oliveira, E. F., & Silveira, M. S. (2018a). Open government data in Brazil a systematic review of its uses and issues. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 1–9. https://doi.org/10.1145/3209281.3209339

de Oliveira, E. F., & Silveira, M. S. (2018b). Open government data in Brazil a systematic review of its uses and issues. *Proceedings Ofthe 19th Annual International Conference on Digital Government Research*, 1–9. https://doi.org/10.1145/3209281.3209339

de Souza, M. F., Tisler, T. R., Castro, G. S. A., & Oliveira, A. L. R. de. (2023). Port regionalization for agricultural commodities: Mapping exporting port hinterlands. *Journal of Transport Geography*, *106*, 103506. https://doi.org/10.1016/j.jtrangeo.2022.103506

Di Ciommo, M. (2015). *Transparency in Brazil: The domestic environment for transparency, access to information and open data*. Development Initiatives. https://devinit.org/wp-content/uploads/2015/03/Transparency-in-Brazil_briefing_March-2015.pdf

Ding, Y., Duan, H., Xie, M., Mao, R., Wang, J., & Zhang, W. (2022). Carbon emissions and mitigation potentials of 5G base station in China. *Resources, Conservation and Recycling*, *182*, 106339. https://doi.org/10.1016/j.resconrec.2022.106339

Estache, A. (2003). *On Latin America's Infrastructure Privatization and its Distributional Effects* (SSRN Scholarly Paper No. 411942). https://doi.org/10.2139/ssrn.411942

Fearnside, P. M. (1999). Social Impacts of Brazil's Tucuruí Dam. *Environmental Management*, *24*(4), 483–495. https://doi.org/10.1007/s002679900248

Flecker, A. S., Shi, Q., Almeida, R. M., Angarita, H., Gomes-Selman, J. M., García-Villacorta, R., Sethi, S. A., Thomas, S. A., Poff, N. L., Forsberg, B. R., Heilpern, S. A., Hamilton, S. K., Abad, J. D., Anderson, E. P., Barros, N., Bernal, I. C., Bernstein, R., Cañas, C. M., Dangles, O., … Gomes, C. P. (2022). Reducing adverse impacts of Amazon hydropower expansion. *Science*, *375*(6582), 753–760. https://doi.org/10.1126/science.abj4017

Gauthier, C., & Moran, E. F. (2018). Public policy implementation and basic sanitation issues associated with hydroelectric projects in the Brazilian Amazon: Altamira and the Belo Monte dam. *Geoforum*, *97*, 10–21. https://doi.org/10.1016/j.geoforum.2018.10.001

595 Hunter, S. B., zu Ermgassen, S. O. S. E., Downey, H., Griffiths, R. A., & Howe, C. (2021).
596 Evidence shortfalls in the recommendations and guidance underpinning ecological
597 mitigation for infrastructure developments. *Ecological Solutions and Evidence*, *2*(3),
598 e12089. https://doi.org/10.1002/2688-8319.12089
599 Hyde, J. L., Bohlman, S. A., & Valle, D. (2018). Transmission lines are an under-acknowledged
600 conservation threat to the Brazilian Amazon. *Biological Conservation*, *228*, 343–356.
601 https://doi.org/10.1016/j.biocon.2018.10.027
602 Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths
603 of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268.
604 https://doi.org/10.1080/10580530.2012.716740
605 Johansson, T. B., & Goldemberg, J. (2002). *Energy for sustainable development: A policy*
606 *agenda*.
607 Joppa, L. N., O'Connor, B., Visconti, P., Smith, C., Geldmann, J., Hoffmann, M., Watson, J. E.
608 M., Butchart, S. H. M., Virah-Sawmy, M., Halpern, B. S., Ahmed, S. E., Balmford, A.,
609 Sutherland, W. J., Harfoot, M., Hilton-Taylor, C., Foden, W., Minin, E. D., Pagad, S.,
610 Genovesi, P., … Burgess, N. D. (2016). Filling in biodiversity threat gaps. *Science*,
611 *352*(6284), 416–418. https://doi.org/10.1126/science.aaf3565
612 Kaufmann, D., & Bellver, A. (2005). *Transparenting Transparency: Initial Empirics and Policy*
613 *Applications* (SSRN Scholarly Paper No. 808664). https://doi.org/10.2139/ssrn.808664
614 Klarenberg, G., Muñoz-Carpena, R., Perz, S., Baraloto, C., Marsik, M., Southworth, J., & Zhu,
615 L. (2019). A spatiotemporal natural-human database to evaluate road development
616 impacts in an Amazon trinational frontier. *Scientific Data*, *6*(1), 93.
617 https://doi.org/10.1038/s41597-019-0093-7

618 Latrubesse, E. M., Arima, E. Y., Dunne, T., Park, E., Baker, V. R., d'Horta, F. M., Wight, C.,
619 Wittmann, F., Zuanon, J., Baker, P. A., Ribas, C. C., Norgaard, R. B., Filizola, N., Ansar,
620 A., Flyvbjerg, B., & Stevaux, J. C. (2017). Damming the rivers of the Amazon basin.
621 *Nature*, *546*(7658), 363–369. https://doi.org/10.1038/nature22333
622 Laurance, W. F. (2018). Conservation and the Global Infrastructure Tsunami: Disclose, Debate,
623 Delay! *Trends in Ecology & Evolution*, *33*(8), 568–571.
624 https://doi.org/10.1016/j.tree.2018.05.007
625 Laurance, W. F., & Arrea, I. B. (2017). Roads to riches or ruin? *Science*, *358*(6362), 442–444.
626 https://doi.org/10.1126/science.aao0312
627 Laurance, W. F., Peletier-Jellema, A., Geenen, B., Koster, H., Verweij, P., Van Dijck, P.,
628 Lovejoy, T. E., Schleicher, J., & Van Kuijk, M. (2015). Reducing the global
629 environmental impacts of rapid infrastructure expansion. *Current Biology*, *25*(7), R259–
630 R262. https://doi.org/10.1016/j.cub.2015.02.050
631 Lupinetti-Cunha, A., Cirino, D. W., Vale, M. M., & Freitas, S. R. (2022). Roadless areas in
632 Brazil: Land cover, land use, and conservation status. *Regional Environmental Change*,
633 *22*(3), 96. https://doi.org/10.1007/s10113-022-01953-9
634 Martínez, D. M., & Ebenhack, B. W. (2008). Understanding the role of energy consumption in
635 human development through the use of saturation phenomena. *Energy Policy*, *36*(4),
636 1430–1435. https://doi.org/10.1016/j.enpol.2007.12.016
637 Menezes, J. A., Confalonieri, U., Madureira, A. P., Duval, I. de B., Santos, R. B. dos, &
638 Margonari, C. (2018). Mapping human vulnerability to climate change in the Brazilian
639 Amazon: The construction of a municipal vulnerability index. *PLOS ONE*, *13*(2),
640 e0190808. https://doi.org/10.1371/journal.pone.0190808

Mere-Roncal, C., Cardoso Carrero, G., Chavez, A. B., Almeyda Zambrano, A. M., Loiselle, B., Veluk Gutierrez, F., Luna-Celino, V., Arteaga, M., Schmitz Bongiolo, E., Segura Tomasi, A., Van Damme, P. A., Lizarro Zapata, D. E., & Broadbent, E. N. (2021). Participatory Mapping for Strengthening Environmental Governance on Socio-Ecological Impacts of Infrastructure in the Amazon: Lessons to Improve Tools and Strategies. *Sustainability*, *13*(24), Article 24. https://doi.org/10.3390/su132414048

Moran, E. F., Lopez, M. C., Moore, N., Müller, N., & Hyndman, D. W. (2018). Sustainable hydropower in the 21st century. *Proceedings of the National Academy of Sciences*, *115*(47), 11891–11898. https://doi.org/10.1073/pnas.1809426115

Nepstad, D., Carvalho, G., Barros, A. C., Alencar, A., Capobianco, J. P., Bishop, J., Moutinho, P., Lefebvre, P., Silva, U. L., & Prins, E. (2001). Road paving, fire regime feedbacks, and the future of Amazon forests. *Forest Ecology and Management*, *154*(3), 395–407. https://doi.org/10.1016/S0378-1127(01)00511-4

Nickerson, S., Chen, G., Fearnside, P. M., Allan, C. J., Hu, T., Carvalho, L. M. T. de, & Zhao, K. (2022). Forest loss is significantly higher near clustered small dams than single large dams per megawatt of hydroelectricity installed in the Brazilian Amazon. *Environmental Research Letters*, *17*(8), 084026. https://doi.org/10.1088/1748-9326/ac8236

Open Data Charter. (2015). *International Open Data Charter*. Open Data Charter. https://opendatacharter.org/wp-content/uploads/2023/12/opendatacharter-charter_F.pdf

Pereira, R. (2021). Public participation, indigenous peoples' land rights and major infrastructure projects in the Amazon: The case for a human rights assessment framework. *Review of European, Comparative & International Environmental Law*, *30*(2), 184–196. https://doi.org/10.1111/reel.12400

664 Pfaff, A., Robalino, J., Reis, E. J., Walker, R., Perz, S., Laurance, W., Bohrer, C., Aldrich, S.,
665 Arima, E., Caldas, M., & Kirby, K. R. (2018). Roads & SDGs, tradeoffs and synergies:
666 Learning from Brazil's Amazon in distinguishing frontiers. *Economics*, *12*(1).
667 https://doi.org/10.5018/economics-ejournal.ja.2018-11
668 Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*,
669 *45*(4), 211–218. https://doi.org/10.1145/505248.506010
670 Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and Opportunites of
671 Open Data in Ecology. *Science*, *331*(February), 703–705.
672 https://doi.org/10.1126/science.1197962
673 Relly, J. E. (2010). A Study of E-government and Political Indicators in Developing Nations
674 with and Without Access-to-Information Laws. In C. G. Reddick (Ed.), *Comparative E-*
675 *Government* (pp. 525–542). Springer. https://doi.org/10.1007/978-1-4419-6536-3_27
676 Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public Data Archiving in
677 Ecology and Evolution: How Well Are We Doing? *PLOS Biology*, *13*(11), e1002295.
678 https://doi.org/10.1371/journal.pbio.1002295
679 Ruijer, E., Grimmelikhuijsen, S., & Meijer, A. (2017). Open data for democracy: Developing a
680 theoretical framework for open data use. *Government Information Quarterly*, *34*(1), 45–
681 52. https://doi.org/10.1016/j.giq.2017.01.001
682 Slough, T., Urpelainen, J., & Yang, J. (2015). Light for all? Evaluating Brazil's rural
683 electrification progress, 2000–2010. *Energy Policy*, *86*, 315–327.
684 https://doi.org/10.1016/j.enpol.2015.07.001
685 Souza, C. M., Kirchhoff, F. T., Oliveira, B. C., Ribeiro, J. G., & Sales, M. H. (2019). Long-Term
686 Annual Surface Water Change in the Brazilian Amazon Biome: Potential Links with

687 Deforestation, Infrastructure Development and Climate Change. *Water*, *11*(3), Article 3.
688 https://doi.org/10.3390/w11030566
689 Souza, M. F. de, Tisler, T. R., Castro, G. S. A., & Oliveira, A. L. R. de. (2023). Port
690 regionalization for agricultural commodities: Mapping exporting port hinterlands.
691 *Journal of Transport Geography*, *106*, 103506.
692 https://doi.org/10.1016/j.jtrangeo.2022.103506
693 Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information
694 quality assessment. *Journal of the American Society for Information Science and*
695 *Technology*, *58*(12), 1720–1733. https://doi.org/10.1002/asi.20652
696 Swanson, A. C., & Bohlman, S. (2021). Cumulative Impacts of Land Cover Change and Dams
697 on the Land–Water Interface of the Tocantins River. *Frontiers in Environmental Science*,
698 *9*. https://www.frontiersin.org/articles/10.3389/fenvs.2021.662904
699 Tisler, T. R., Teixeira, F. Z., & Nóbrega, R. A. A. (2022). Conservation opportunities and
700 challenges in Brazil's roadless and railroad-less areas. *Science Advances*, *8*(9), eabi5548.
701 https://doi.org/10.1126/sciadv.abi5548
702 Transparency and Accountability Initiative. (2014). *Open Government Guide*. The Transparency
703 and Accountability Initiative. https://www.opengovpartnership.org/wp-
704 content/uploads/2019/05/open-gov-guide_summary_all-topics1.pdf
705 Tucker Lima, J. M., Valle, D., Moretto, E. M., Pulice, S. M. P., Zuca, N. L., Roquetti, D. R.,
706 Beduschi, L. E. C., Praia, A. S., Okamoto, C. P. F., da Silva Carvalhaes, V. L., Branco, E.
707 A., Barbezani, B., Labandera, E., Timpe, K., & Kaplan, D. (2016). A social-ecological
708 database to advance research on infrastructure development impacts in the Brazilian
709 Amazon. *Scientific Data*, *3*(1), 160071. https://doi.org/10.1038/sdata.2016.71

710 Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open
711 data quality measurement framework: Definition and application to Open Government
712 Data. *Government Information Quarterly*, *33*(2), 325–337.
713 https://doi.org/10.1016/j.giq.2016.02.001
714 Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert,
715 K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The Availability of Research
716 Data Declines Rapidly with Article Age. *Current Biology*, *24*(1), 94–97.
717 https://doi.org/10.1016/j.cub.2013.11.014
718 Wang, H.-J., & Lo, J. (2016). Adoption of open government data among government agencies.
719 *Government Information Quarterly*, *33*(1), 80–88.
720 https://doi.org/10.1016/j.giq.2015.11.004

721 **Appendix A: Supplemental Figures**

Figure 1

Column chart of uses of infrastructure data attributes for each paper in the literature review.

The proportion of times each attribute was used per paper in the literature review conducted in the Web of Science database (WOS) for the 2011-2018 period (orange), and the proportion of times each attribute was ranked in the top five (blue) for importance to include in an infrastructure data set by survey participants. EIA stands for environmental impact assessment. The dashed line at 0.4 is the cutoff for the attributes identified as critical for infrastructure datasets.

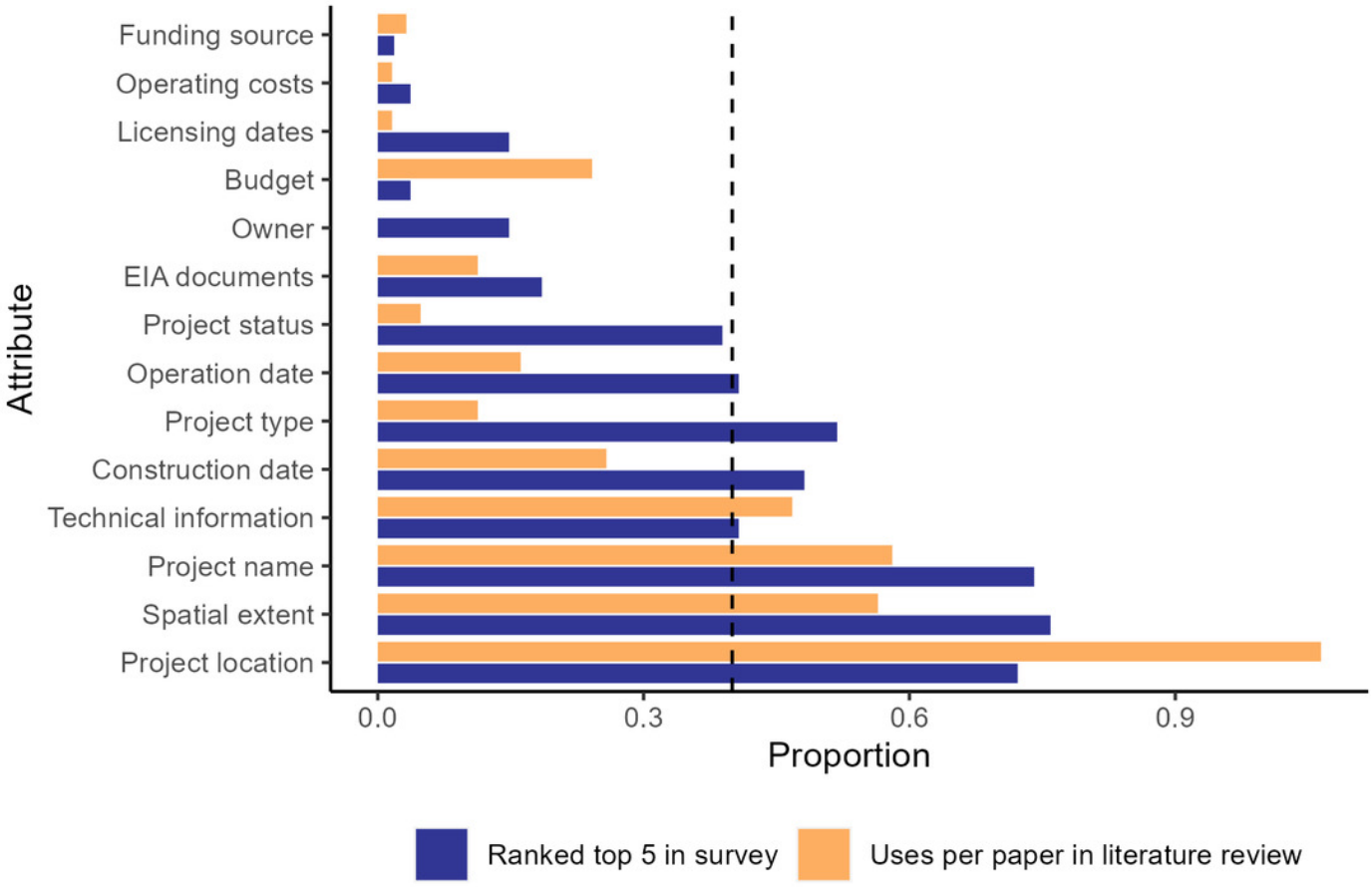

Figure 2

Column chart displaying number of times survey participants search for and found infrastructure data sets from various sources.

Number of infrastructure data sets that survey participants searched for and found from the data sources in which they originally searched (blue), searched for and found from a different search data source (orange), or searched for but did not find in any data repository (red). Participants may have searched for the same data set from different sources. For example, a participant may have searched for a data source from the government, not found it (yellow bar for government data source), then searched in academic data sources and successfully found the data set there (blue bar for academic data source), but we were unable to track this information.

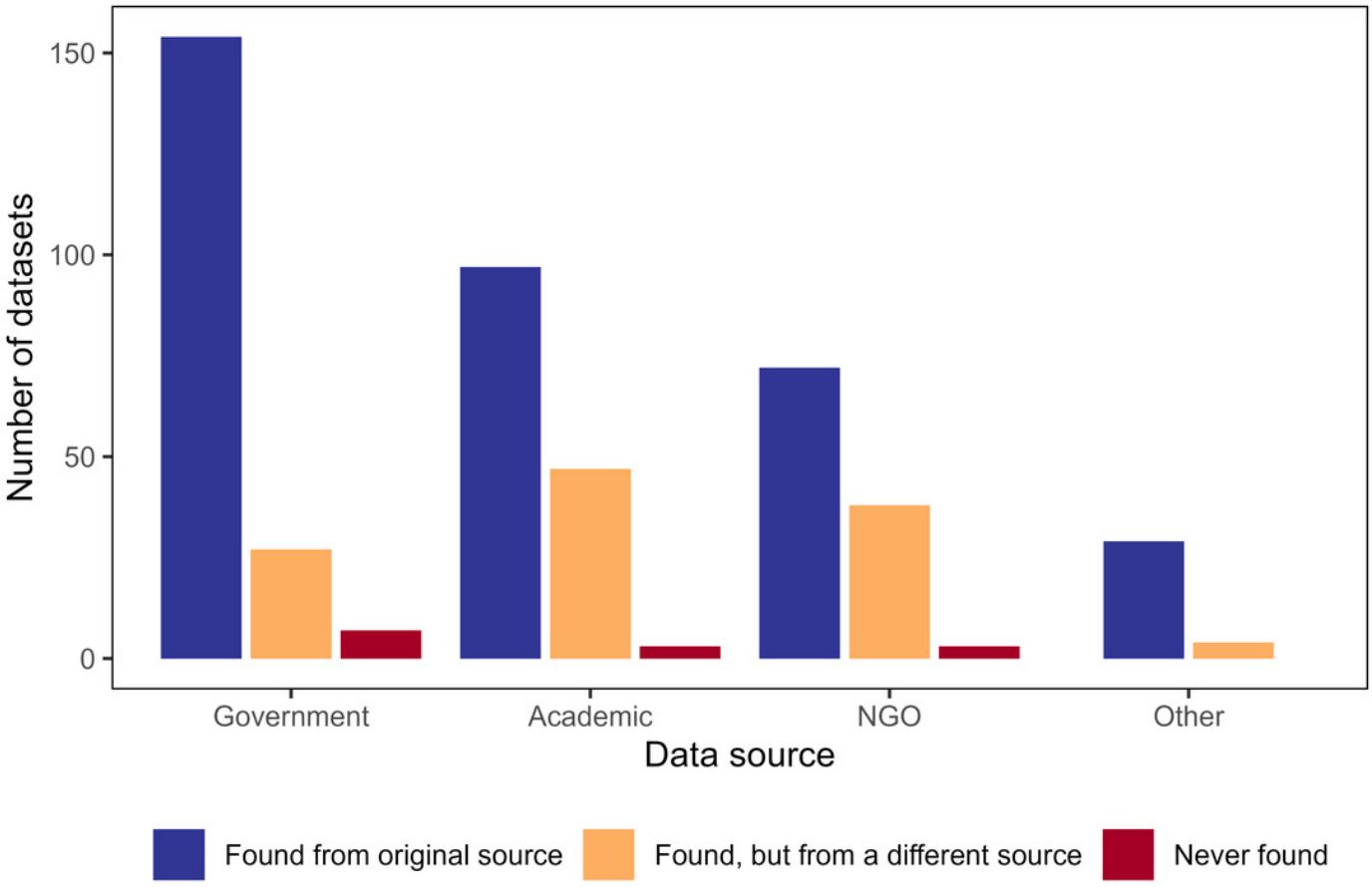
150
100
50
0
Number of datasets
Government
Academic
NGO
Other
Data source
Found from original source
Found, but from a different source
Never found

Figure 3

Survey responses to questions about data quality.

Responses to questions about data quality for: (A) data accuracy (both spatial and non-spatial components); (B) task-independent standards – completeness, currency (< 1 year since update), machine readability, and metadata quality. N/A value for metadata quality indicates respondent did not require metadata for the dataset they used.

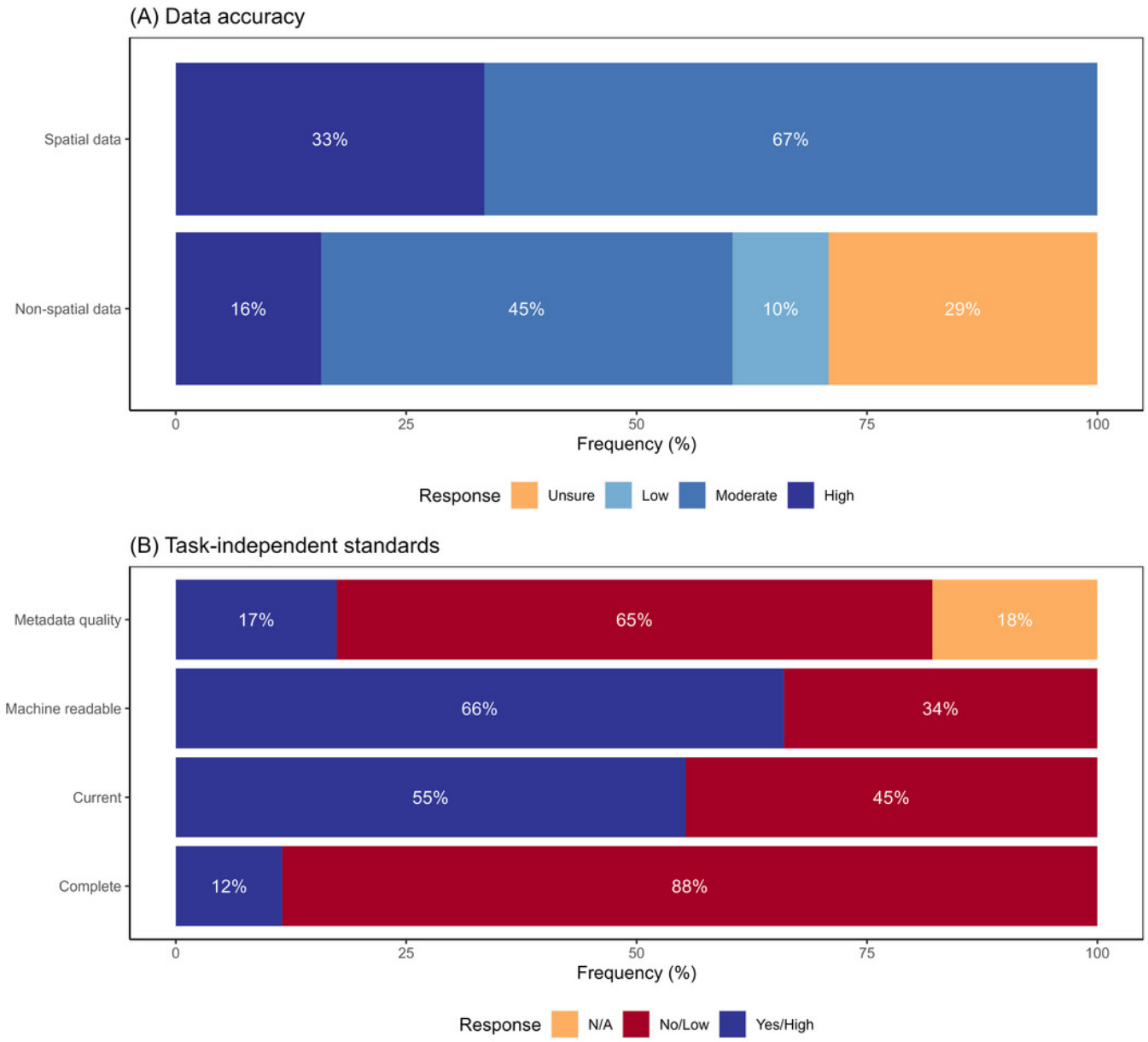
(A) Data accuracy
Spatial data
Non-spatial data
33%
67%
16%
45%
10%
29%
0
25
50
75
100
Frequency (%)
Response
Unsure
Low
Moderate
High
(B) Task-independent standards
Metadata quality
Machine readable
Current
Complete
17%
65%
18%
66%
34%
55%
45%
12%
88%
0
25
50
75
100
Frequency (%)
Response
N/A
No/Low
Yes/High

Figure 4

Time spent searching for and formatting data about large infrastructure projects in the Brazilian Amazon.

Time spent searching for and formatting data about large infrastructure projects in the Brazilian Amazon.

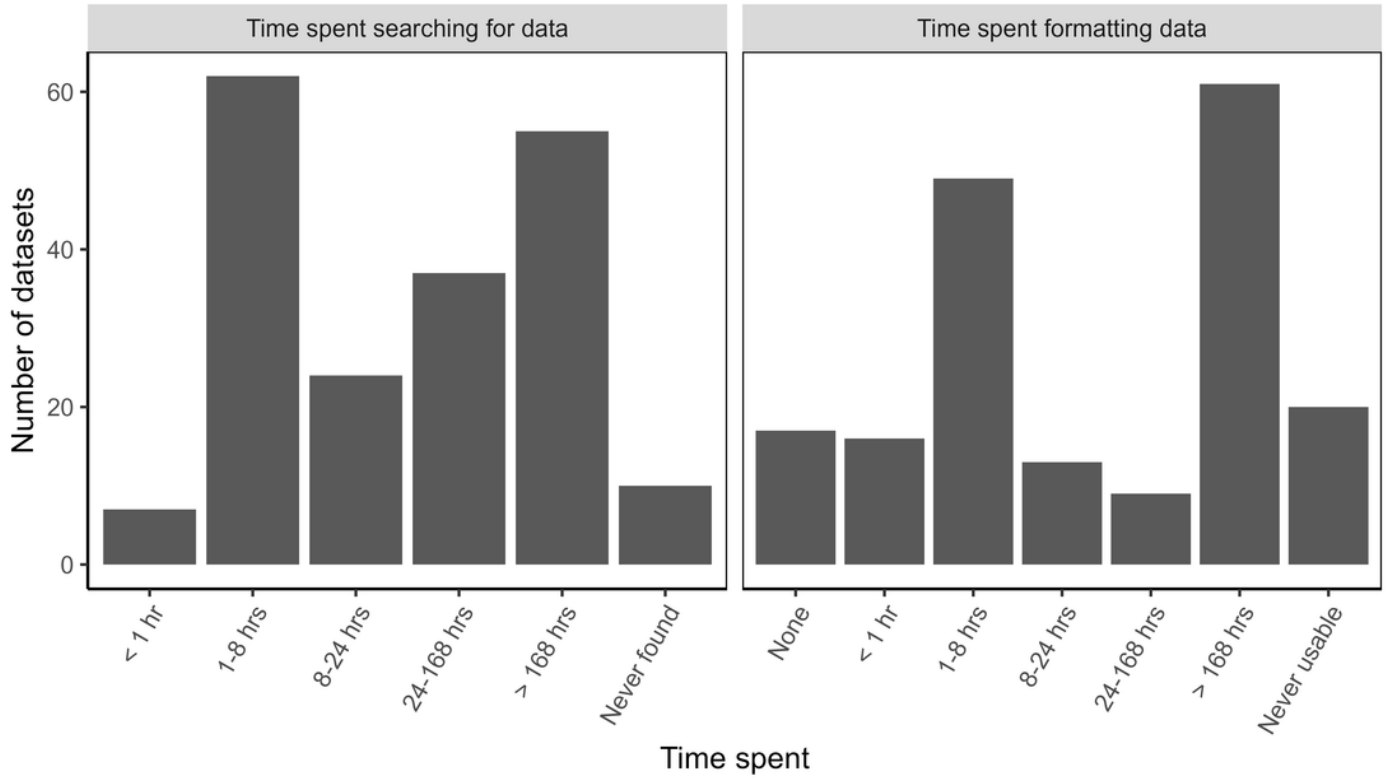

Figure 5

Research themes, infrastructure types, and attributed used according to survey respondents.

Number of articles in the systematic literature review conducted in the Web of Science database (WOS) for the 2011-2018 period grouped by (A) thematic research area of the articles (LULC stands for land use/land cover); (B) infrastructure project type researched; and (C) data attributes used in the research (EIA stands for environmental impact assessment).

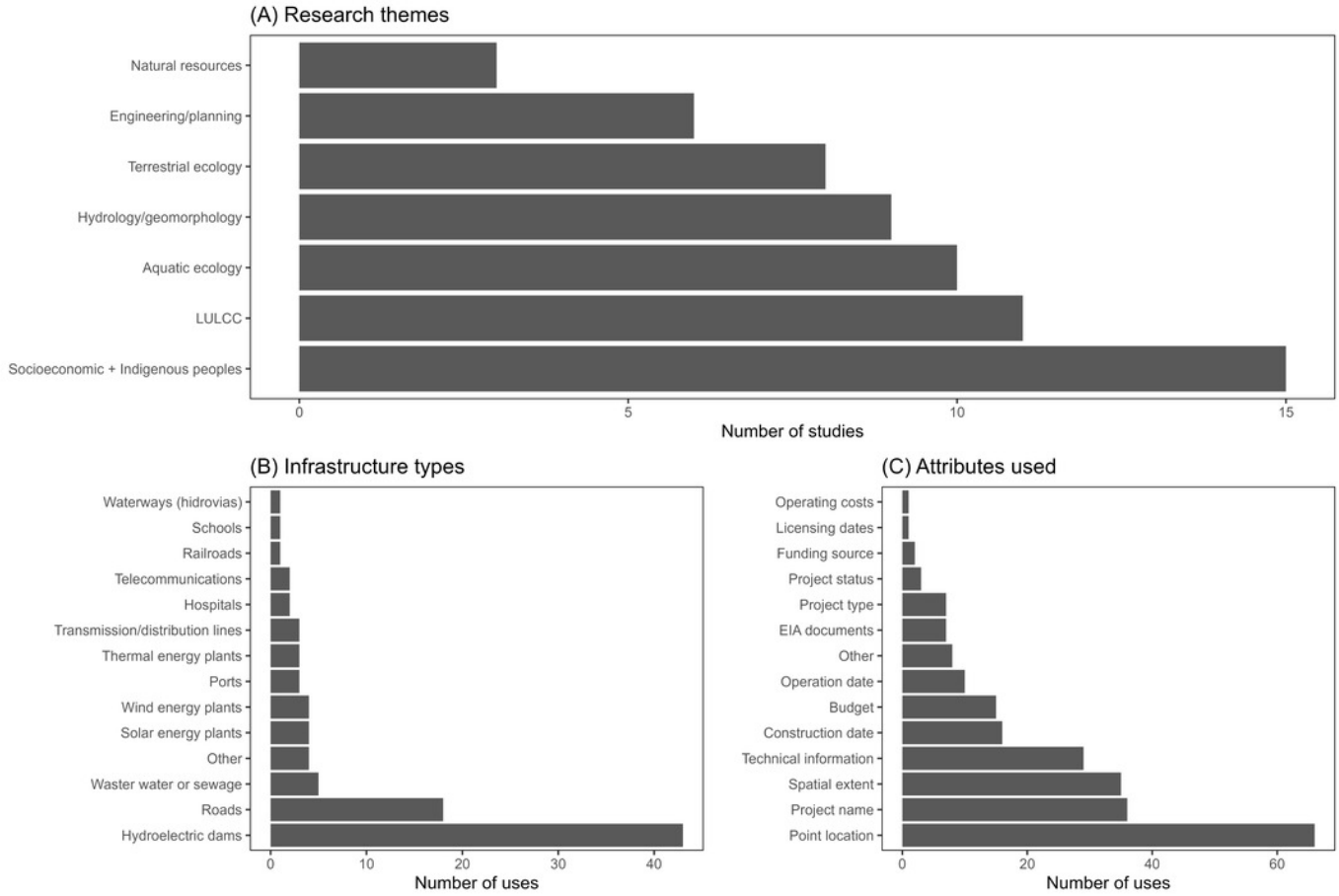

Figure 6

Demographic data of survey respondents.

Survey responses for A) participants' primary work locations and B) career type; C) themes of the projects for which participants needed infrastructure data. "LULCC" stands for land-use/land-cover change; D) types of infrastructure data searched for.

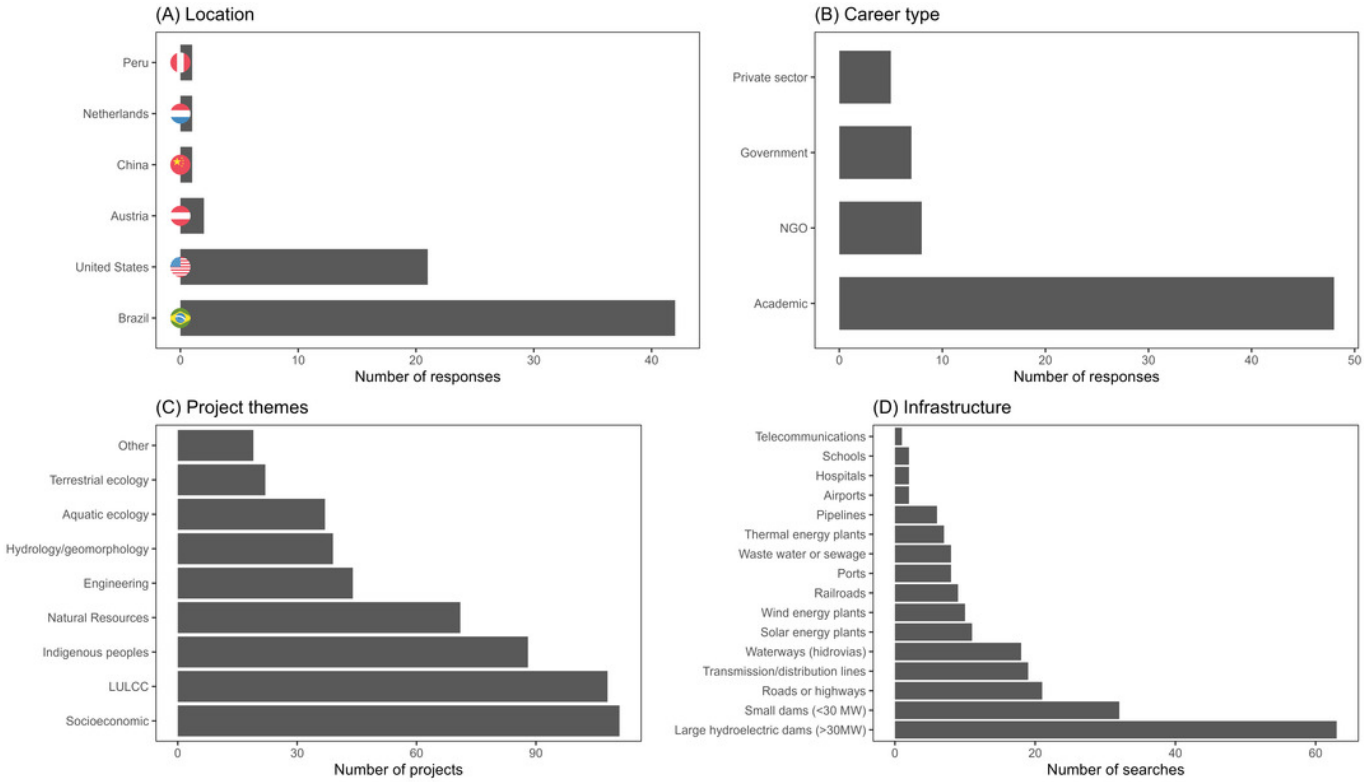

Figure 7

Proportion of searches by survey respondents for infrastructure project attributes categorized by project type.

Proportion of searches by survey respondents for project attributes categorized by project type. Projects that had fewer than 10 attribute searches are not shown. Attributes are ordered according to their overall popularity from most to least frequently searched (left to right). Data were grouped into energy distribution (pipelines, transmission/distribution lines), energy generation (large and small dams, solar, thermal, wind), sewage (wastewater and sewage), and transportation (ports, railroads, roads and highways, and waterways/hidrovias).

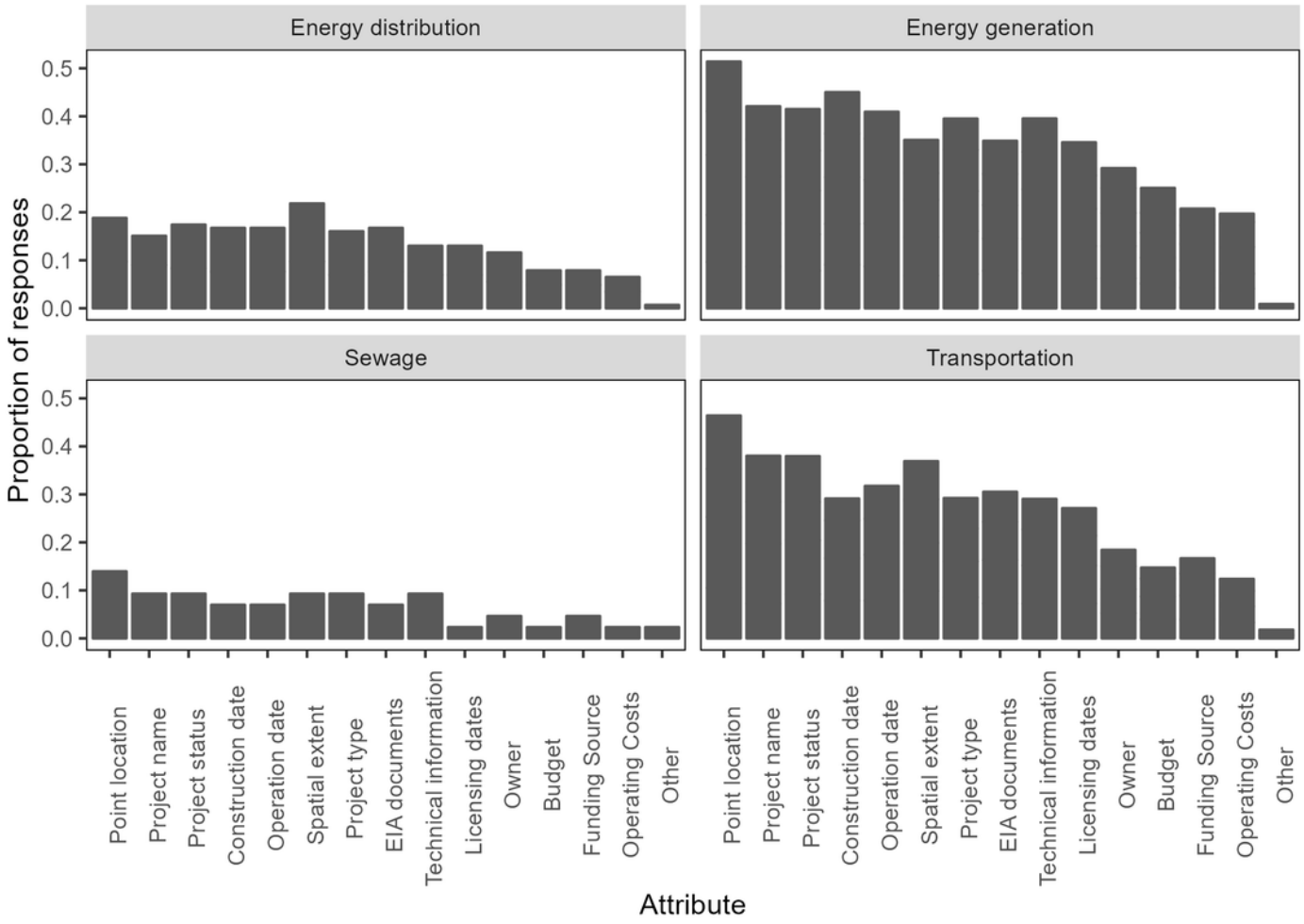

Figure 8

Responses to questions about data quality for datasets from government and non-government sources.

Responses to questions about data quality for datasets from government and non-government sources: (A) data accuracy (both spatial and non-spatial components); (B) task-independent standards – completeness, currency (< 1 year since update), machine readability, and metadata quality. N/A value for metadata quality indicates respondent did not require metadata for the dataset they used

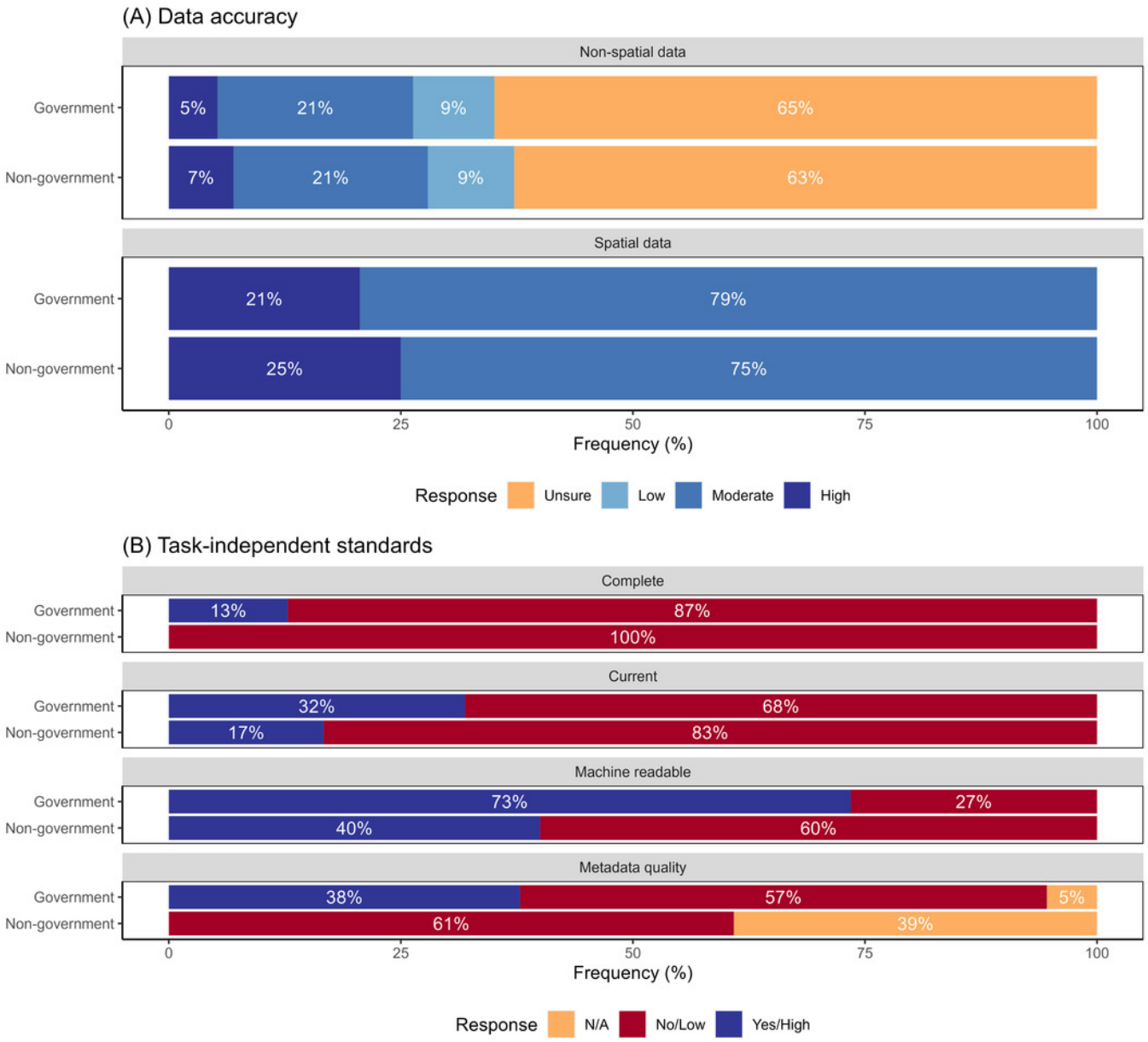
(A) Data accuracy
Non-spatial data
Government
Non-government
5%
21%
9%
65%
7%
21%
9%
63%
Spatial data
Government
Non-government
21%
79%
25%
75%
0
25
50
75
100
Frequency (%)
Response
Unsure
Low
Moderate
High
(B) Task-independent standards
Complete
Government
Non-government
13%
87%
100%
Current
Government
Non-government
32%
68%
17%
83%
Machine readable
Government
Non-government
73%
27%
40%
60%
Metadata quality
Government
Non-government
38%
57%
5%
61%
39%
0
25
50
75
100
Frequency (%)
Response
N/A
No/Low
Yes/High

Figure 9

Time spent searching for and formatting data for large infrastructure projects in the Brazilian Amazon, divided by source of data.

Time spent searching for and formatting data about large infrastructure projects in the Brazilian Amazon. Note that “all sources” represents the total number of datasets searched for or formatted and includes datasets found only from government or non-government sources and datasets found from both government and non-government sources. As a result, number of data sets from all sources (blue) can exceed the sum of data sets from government only (red) and non-governmental only sources (yellow).

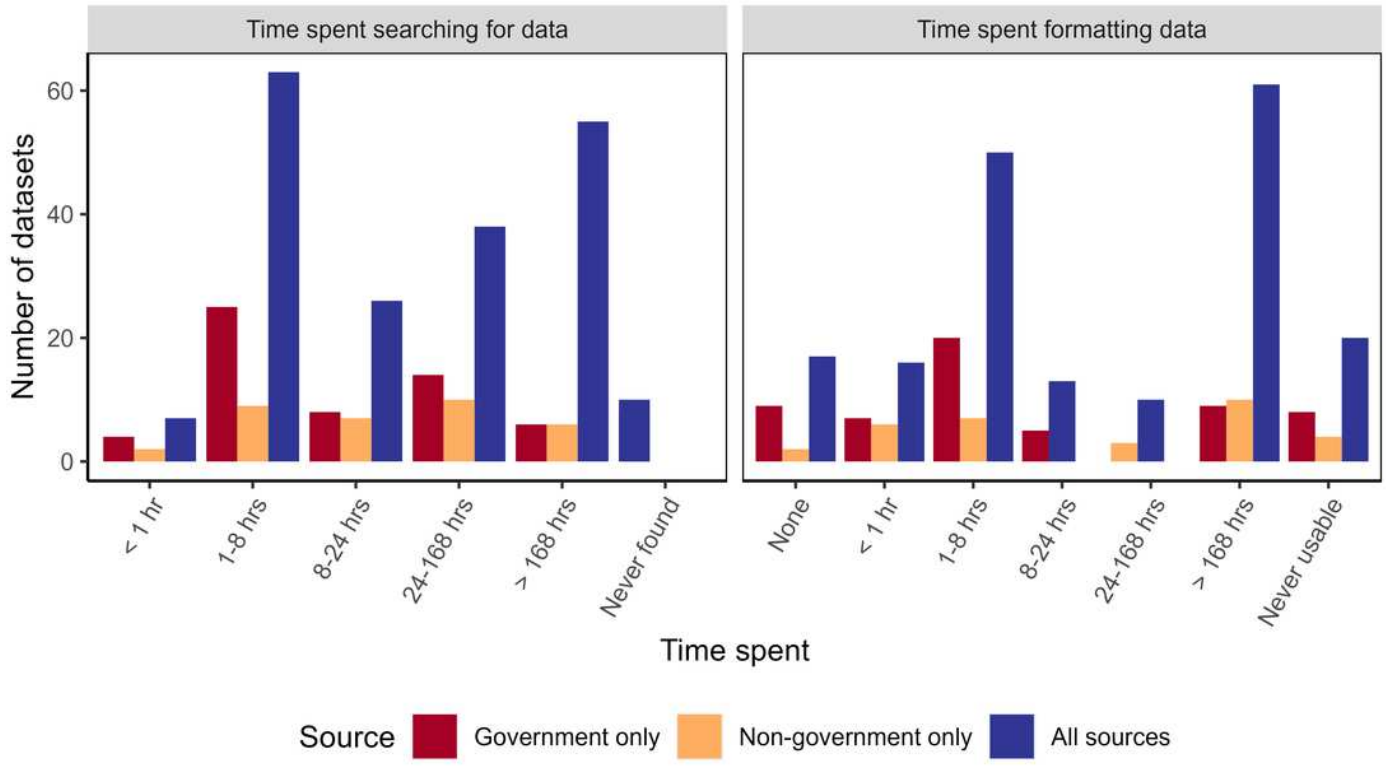

Table 1(on next page)

The quality and content of publicly available infrastructure datasets.

The quality and content of publicly available infrastructure datasets based on our proposed critical attributes and task-independent standards.

	Federal roads	Large dams
Dataset origin	Departamento Nacional de Infraestrutura de Transportes	Agencia Nacional de Energia Elétrica
Critical attributes		
Project Name	Yes	Yes
Geographic extent	Yes	No (point location only)
Basic technical info	No	Yes (capacity, drainage area)
Construction date	No	No
Project type	Yes (pavement status, federal status, road size)	No
Operation date	No	No
Task-independent standards		
Spatially accuracy	98%	96%
Completeness (2 columns)	100 %	94.3%
Currentness	No information	Yes (updated 5/8/19)
Machine readable	Yes (shapefile, KMZ)	Yes (shapefile, KMZ)
Metadata		
Author	No	No
Date of creation/update	No	Yes
Attribute descriptions	No	No
Datum	No	No