This is a preprint of an article published in Library & Information Science Research: Stvilia, B. & Gibradze, L. (2022). Seeking and sharing datasets in an online community of data enthusiasts. *Library & Information Science Research, 44*(3), 101160. https://doi.org/10.1016/j.lisr.2022.101160

# Seeking and sharing datasets in an online community of data enthusiasts

Besiki Stvilia[1], Leila Gibradze

College of Communication and Information, Florida State University, 142 Collegiate Loop, Tallahassee, FL 32306-2100, USA

## ABSTRACT

This study examined discussions of the r/Datasets community on Reddit. It identified three activities in which the community engaged: question answering, data sharing, and community building. Members of the community used 21 types of data and information sources in their activities. The findings of this research enhance our understanding of the activity structures, data and information sources used, and challenges and problems encountered when users search for, share, and make sense of datasets on the web, outside the traditional information and data ecosystems. Data librarians and curators can use the findings of this study in the design of their data management and reference services. The typology of data sources and the metadata model developed through this study can be used in annotating and categorizing data sources and informing the design of descriptive metadata schemas and vocabularies for datasets.

## 1. Introduction

The rapidly decreasing cost of hardware, the automated collection, processing, and sharing of large-scale data, and the maturity of machine-learning and data-mining algorithms and models that enable their wider application with increased accuracy have generated a surge in the number of large-scale datasets and the use of data-centric computational approaches in almost every area of our lives (Boyd & Crawford, 2012). The increased dissemination of data by governments, universities, academic societies, publishers, Internet consumer companies, and other businesses has dramatically increased the public's access to data (Demchenko et al., 2012). This deluge of data has created opportunities ranging from data-driven marketing and personalized consumer services to data-driven open science and the use of external datasets in teaching data science and computational courses. This flood of data and its ease of collection also have significant consequences for different aspects of our individual lives and for society in general regarding privacy, information credibility, and security (Boyd & Crawford, 2012).

The increased availability of open datasets also creates data reuse and integration opportunities to deepen understanding of a particular phenomenon, identify latent relations across different contexts, and foster data-centric product and service innovations. People reuse and

[1] Corresponding author. E-mail address: bstvilia@fsu.edu (B. Stvilia).

repurpose existing datasets more often instead of creating original data for their projects (Groves, 2011). In addition, to develop deeper, more accurate, and complete insights into a problem and identify and make connections among different issues and problems, researchers need to capture and include the context of the problem in their examinations. Because researchers and businesses often cannot generate data that captures that context, they may have to rely on reusing third-party, external datasets and ontologies (Coletta et al., 2012). Furthermore, secondary research practices and research replication rely on reusable data (National Academies of Sciences, Engineering, and Medicine, 2019). However, sources of third-party data may vary. Similar to users seeking and using information and knowledge from both traditional information sources and the web, users (including students and faculty) search for and use data from both traditional curated repositories and uncurated sources on the web.

## 2. Problem statement

A significant body of literature exists on research data curation in scholarly data repositories and academic libraries (e.g., Cragin et al., 2010; Lee & Stvilia, 2017; Tenopir et al., 2014). Likewise, an abundance of research has been produced on data practices in different academic research communities and communities of citizen scientists (e.g., Birnholtz & Bietz, 2003; Borgman et al., 2007; Ferguson et al., 2014; Stvilia et al., 2015).

In addition, literature on social questioning and answering (Q&A) and content sharing in online communities is extensive (e.g., Adamic et al., 2008; Fu, 2019; Oh, et al., 2008; Stvilia et al., 2008). Yet few research studies have been conducted on data-seeking, data-sharing, and data-reuse practices for data on the web. In particular, research is lacking on how people seek and share information on datasets on online social Q&A and discussion sites. This study contributes to addressing this gap by examining the following research questions:

1. What are the major activities of online social Q&A and discussion communities that are focused on datasets?
2. What tools and sources do members use in their activities?
3. What problems and challenges do members encounter in those activities?

## 3. Literature review

What are considered data varies from one discipline to another. For example, to a materials scientist, physical samples of material are data, whereas to an experimental physicist, a graph published in a scholarly paper can be data (Stvilia et al., 2015). Likewise, different entities can generate data, such as government agencies, individual researchers, research teams, laboratories, research centers, businesses, and international organizations. Digital data may range from raw structured data streamed from sensors or generated by computer simulations to text documents, laboratory notes, website content, diagrams, graphs, and software code (Borgman et al., 2007; Stvilia et al., 2015). Another type of data is metadata, which refers to structured data that enables specific functionalities. Metadata is essential for discovering, making sense of, and reusing data (Lee & Stvilia, 2017).

Multiple general models of information seeking have been proposed in the information behavior literature. These models conceptualize the structure of an information-seeking activity as the interactions among the user's information need, context, goal-oriented actions, and tools (e.g., Wilson, 1997, 2006). One type of collaborative information seeking and sharing is social Q&A. Studies of social Q&A have examined the need and motivations for asking and answering questions and the types of sources used to answer questions (Cunningham & Hinze, 2014; Fu,

2019, Oh et al., 2008). Another group of social Q&A studies has proposed predictive models for evaluating the quality of questions and answers and for identifying experts (e.g., Fu, Wu, & Oh, 2015; Pal, Harper et al., 2012). Prior studies have examined the types of information sources used in social Q&A on different topics. However, examinations of the social Q&A and discussion groups of data enthusiasts and practitioners are lacking, including their information behaviors and the structures of those behaviors, as well as the sources and tools used.

Another relevant literature genre that informs this study is the digital data curation literature (e.g., Higgins, 2008; Lee & Stvilia, 2017). Although general infrastructure components of digital data curation are shared across different disciplines, the research project tasks, the types of data and digital objects produced, and the norms followed in managing and sharing data may vary (Borgman et al., 2007; Chen & Chen, 2020). Work and research projects may involve seeking, collecting, and aggregating different kinds of data. For instance, in health care, the integration of patients' biomedical data, including genomic data and data collected from the medical Internet of Things, with electronic medical records and pharmacy data can enable personalized or precision healthcare (Dash et al., 2019). There is little research, however, on how users seek, select, and obtain relevant datasets for their research projects on the web.

One of the main inhibitors of data sharing and reuse is individuals' concern about the quality of data (Stvilia et al., 2015; Wu & Worrall, 2019). Quality is defined as "fitness for use" (Juran, 1992), and data quality, along with privacy and access, are critical ethical aspects of data use. The quality of data determines the quality of the research findings, teaching, business decisions, and policies. Ultimately, it also affects human lives (Mason, 1986; Stvilia et al., 2007). To support data-driven innovation, knowledge creation, and policy making, data need to be reusable. Data owners may be concerned about the quality or documentation of their data and its potential misuse or misinterpretation by others (Stvilia et al., 2015; Wu & Worrall, 2019). The users, on the other hand, need useful and high-quality data, not just big data (Ng, 2021). Data creators usually collect or assemble datasets for specific purposes or uses. If data are not properly documented, understanding those purposes is often a challenge and a barrier to data reuse (Swarup et al., 2018). Researchers have conceptualized research data quality and studies of scientists' perceptions of and priorities for data quality and their data quality assurance skills (e.g., Gutmann et al., 2004; Huang et al., 2012; Stvilia et al., 2015). Furthermore, several general quality assurance standards and approaches have been used in the industry (e.g., Six Sigma, ISO 9000). The literature also includes data reusability frameworks, such as the FAIR, which comprises the criteria of data findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016). However, empirical studies that focus on dataset quality evaluation practices on the web, including on social Q&A and discussion sites are lacking.

## 4. Research design

To address the research questions above, the study analyzed discussion threads from the r/Datasets subreddit. Reddit is one of the 10 most highly used social media platforms. It has millions of daily active users and more than a hundred thousand active interest- or topic-based discussion communities (i.e., subreddits; Auxier & Anderson, 2021; Clement, 2021). The study examined a sample of 1,232 posts and 6,813 related comments collected from the r/Datasets subreddit on January 3, 2021. The subreddit was created on October 8, 2009, and at the time of data collection, the r/Datasets subreddit had 131,000 members. Reddit does not report the total number of posts in its subreddits.

The authors used Reddit's application program interface (API) to collect the sample. Only submissions marked as "hot" or "top" by the API were requested. Hot submissions are submissions that have recently received a high number of upvotes or comments, whereas top submissions are those that have received the greatest number of upvotes. The date range of the sample was 2010-2020, and the median number of comments per post was three.

The authors next analyzed the sample by content analysis. For that purpose, the authors combined each post and its comments into a single thread case. The unit of analysis was an individual thread. The content analysis was guided by activity theory (Kaptelinin & Nardi, 2012). Activity theory conceptualizes a general, hierarchical structure of activity that comprises goal-oriented actions mediated by tools and community. In addition, activity theory includes a concept of contradiction. It is defined as a problem or misalignment within a particular activity or between related activities. Identifying and addressing contradictions are important as it can lead to innovation and evolvement of new, more effective activity systems and structures. The study's research questions were already grounded in the general concepts of the activity structure (i.e., activity, tools, contradictions). The authors used the research questions and activity theory to develop a priori codes for the content analysis, identify instances of those codes in the sample, and iteratively categorize and aggregate both a priori and emerging codes according to the hierarchical structure of activity (i.e., the activity-action model).

The authors selected a random subset of 120 threads from the dataset to develop a coding scheme for the content analysis. Each of the authors independently analyzed this subset for both *a priori* and emerging themes. They compared their individual lists of codes, aggregated and mapped them. Twenty themes were identified in the first round of the comparison. The authors inductively aggregated these 20 themes into five general categories that matched the research questions and high-level concepts of activity theory: activity, actions, motivations, tools, and problems (Bailey, 1994). They used these five categories of 20 themes to code the dataset. Each of the authors coded half of the sample independently. They used another random 120 threads from the recoded sample to evaluate the reliability of their coding. The kappa statistic for the intercoder reliability of the subset was 0.85. This score qualifies as a good agreement level (Carletta, 1996). The authors discussed and resolved the cases on which they disagreed, and then updated the code assignments for other similar cases in the complete dataset.

In addition, the study used the labeled latent Dirichlet allocation (LLDA) to generate word profiles for the categories. Specifically, the study applied the LLDA to the data (i.e., 1,232 threads) to construct latent topic models that corresponded to the five general category codes from the manual content analysis. The main advantage of the LLDA is that it can handle multiple topic codes assigned to data cases and generate topic models with greater accuracy than can other supervised learning methods (Ramage et al., 2009). Each profile included the top 100 terms ranked by their probabilities for that topic. The first author adapted Nakatani Shuyo's Python implementation of the LLDA (https://github.com/shuyo/iir/blob/master/lda/llda.py) to generate the profiles.

Finally, the authors extracted the 3,514 URLs contained in the sample. They then selected the 100 most frequently found Internet domains in that set of URLs and categorized them. The authors used the resultant categories to identify the types of information and data sources used by the community in their activities.

The study used publicly available open data in this study. Even so, some members of the Reddit community might not realize that outsiders can use their contributions for research

purposes; hence, the authors modified and paraphrased posts and comments quoted in this paper to obfuscate their connections with specific member handles.

## 5. Findings and discussion

This section of the paper reports findings of the content analysis of 1,232 discussion threads of the r/Datasets subreddit and interprets them through the literature.

### *5.1. Major activities*

Through the content analysis of the sample, the study identified three distinct activities performed in the community: asking and answering questions, disseminating information, and building the community (see Fig. 1). Specifically, members asked and answered questions on how to find and use data. They requested datasets on specific populations, businesses, events, or activities; concrete datasets they had seen or heard about; datasets on a specific problem; datasets of a specific scale or quality; or datasets that were similar to, or complemented, a specific dataset (see Fig. 2). Members also searched for pretrained machine-learning models for transfer learning or state-of-the-art datasets for training or benchmarking a particular machine-learning algorithm. The main purpose for requesting or sharing a dataset or service was its eventual use. Hence, members also asked questions about data use, such as how to use a specific data analysis technique; how to store a big dataset; and how to aggregate, visualize, or cite data. Finally, members asked for advice on how to begin or transition to various data careers. In their responses, members shared how-to guides for repository APIs and data scraping, tutorials, blogs, and data analysis courses.

[ Fig. 1 about here ]

[ Fig. 2 about here ]

In the dissemination activity, members pushed information on datasets, data repositories, and services out to the community. Often, data were not freely available for download but were exposed through HTML pages from company or organizational websites. Members used software such as Beautiful Soup to extract data from those pages and share it with the community. Many of them expressed a sense of pride and accomplishment from being able to make data more accessible and reusable.

They also shared information on the scripts and applications used to scrape from websites, generate synthetic data, and clean or visualize data (see Fig. 3). In addition, members shared data papers and articles that used or described datasets, data bibliographies, lists of data repositories, trained machine-learning models, and data-related news.

[ Fig. 3 about here ]

When sharing datasets, references to repository APIs, and other data products and information services, members often provided descriptive and administrative metadata for those products, such as the list of data fields, license information, information about the data quality and use cases, and the components and configuration of the repository's infrastructure.

In the community-building activity, moderators and members praised other members for sharing useful and unique datasets and data tools. They conducted regular polls and open-ended discussions to identify members' needs for data and assemble bibliographies of datasets and

services. The community also surveyed members about the technologies they used in their data work and asked them to share ideas and hypotheses for a specific dataset (see Fig. 4).

[ Fig. 4 about here ]

*5.2. Major shared action-based themes*

The data Q&A and dissemination activities were closely intertwined. The next subsections describe shared action-based themes found in the analysis of activity conversations.

*5.2.1. Obtaining and aggregating data*

To answer questions, create data, or use data, members often needed to perform traditional tasks of data management, such as discovery, identification, and aggregation. Members requested different kinds of data (e.g., structured, unstructured, text, multimedia) and their access to data could vary. Some datasets were published as open access, whereas other data were accessible only through secondary publications, such as webpages, pdf files, or graphs. To obtain the latter, they might have to engage in complex reverse engineering of data from the secondary publication sources, which could involve scraping, aggregation, and normalization.

Similar to meeting an information need, meeting a data need may require more than one dataset (Bates, 1989). Frequently, in addition to finding and identifying datasets useful to meet the seeker's need, the seeker and the community deliberated how they could aggregate and link related datasets to complete the task. They referenced data aggregation tools, such as crosswalks, mapping tables, and ontologies. Furthermore, the community discussed the ethical and legal aspects of data extraction. It has long been known in the literature that the safe sharing of individual datasets does not necessarily mean that the sharing of their aggregation is ethical or legal also (Mason, 1986). These findings remind us that information technology and data science education programs need to train students not just in methods of data extraction and aggregation, but also in how to conduct those actions in an ethical and legal way.

*5.2.2. Sensemaking*

Seekers' data needs could be ill-structured and not welldefined. Seekers might be aware of the problem or issue they would like to study but still not have a clear design for addressing or examining the problem, including what data they might need or how to obtain it. To meet ill-structured data needs and associated ill-defined questions or questions that might not have ready answers, members often initiated a "reference interview" process to identify the searcher's "conscious need" for data as opposed to the searcher's "formalized" question posted on the subreddit (Taylor, 2015). In those instances, the activity comprised not just social information foraging, but also social sensemaking to build more accurate and complete representations of the searcher's need, the related task requirements, and the data needed to meet those requirements (Pirolli, 2009). The process might comprise sketching out a schema for the data sought, asking the seeker to clarify the purpose of the project, or asking the seeker about the intended use of the dataset.

Similarly, when sharers shared information on datasets or data services, the community often asked them to share the schemas and software codes used to generate or analyze those datasets, or both. The code coupled with the data made a use case. Use cases helped the community learn about possible uses and applications of the data.

Furthermore, data dissemination can be a complex sociotechnical problem. The sharer might be uncertain about or unfamiliar with the data dissemination process, what repositories or infrastructure architectures could meet the needs and preferences of both the sharer and potential users, what format to use, and how data had to be prepared and documented for sharing. Hence, the sharer and community might collaborate to co-construct a representation of the sharing process.

*5.2.3. Collaborating and crowdsourcing*

Requesting or sharing a dataset could trigger a discussion of a more formalized collaboration. Some seekers offered to pay for help to generate the requested data. In other instances, members volunteered their help if the seeker's project matched their interests or was perceived as being of high value to the community or society. Finally, some members would offer assistance with data collection or scraping for a fee.

> *I could build you a scraper that collects all of this data on a regular basis if you like, but it won't be free. I don't have nearly enough time for myself. (t440)*

Like question answering, data sharing too could lead to a conversation on a concrete potential collaboration. A sharer might explicitly solicit and receive help with preparing data for sharing or cleaning a dataset or improving the quality of a script. A sharer might also ask the community to provide feedback on a shared dataset, test a data service or code, evaluate a trained machine-learning model, or evaluate other products of data analysis.

Some members who began or worked for a data science-based business or not-for-profit organization used the event of sharing a dataset or a service to recruit volunteers to crowdsource the data collection for and management of their repositories. Members also approached the community with specific collaboration requests triggered by particular events, such as organizing a hackathon on the COVID-19 pandemic or other causes and problems of high societal impact.

The literature suggests that domain expertise, the diversity of domain expertise, effective work coordination, task routing, and communication mechanisms are important for successful collaborations (Cosley et al., 2007; Stvilia et al., 2008, 2011; Wiggins & Crowston, 2015). Furthermore, member profiles in an information system providing information on members' personalities, resources and skills, reputation, affiliation, and culture could facilitate collaborator selection (Bozeman & Corley, 2004; Stvilia et al., 2017).

*5.2.4. Evaluating data value, cost, and quality*

A seeker's request for a dataset often triggered feedback from the community on the intended use and value of the data or the problem the seeker intended to address with the data. Members also discussed data flaws related to different quality criteria, as well as possible ethical and legal issues related to data creation, aggregation, sharing, and use. The community might comment on the importance of the seeker's data project idea, its feasibility, and how project could be extended or revised. Because many of those projects were data-creation or data-generation projects, the evaluation of their significance was shaped by the uniqueness of the datasets they expected to produce. Having access to a dataset search engine that allowed them to identify related or nearly duplicate datasets, as well as to link project ideas with the datasets they used, would help with this evaluation.

Similarly, sharing a dataset might trigger a discussion of its value. The community's perceived value of a dataset was shaped by its perceived usefulness. To evaluate the usefulness of a dataset, the community might ask the sharer to specify how the data would be used, such as what questions it could answer, what insights it could help develop, or what services it could enable. In

addition, the community often referred to already available datasets or services as benchmarks to gauge the uniqueness or the marginal value of a dataset.

In addition, considering that sharing a large dataset or maintaining a data repository or service can carry a significant cost to the sharer, the sharer might decide to gauge the community's interest in the data before sharing it. Furthermore, members often engaged in assessing the costs of different alternative options for sharing their data to identify the infrastructure options that met their budgets. To help users calculate the cost of big data ownership, including providing public access to it, libraries could develop guides that would list different infrastructure alternatives for big data sharing and their costs.

The third facet of the community's feedback was data quality. The community frequently provided feedback on the quality of a dataset or critiqued the overall idea of a data project. Some members might also replicate a shared project with a different dataset or point to quality problems or errors in the script used to collect the data. For data repositories, it is important that this kind of feedback is not lost and is linked with the evaluated datasets. Data quality evaluation is not free. In this way, a future user of the dataset can access the prior quality evaluations, use a cleaned version of the dataset, or take those evaluations into account when deciding whether to use the data.

*5.3. Tools and sources*

Using and providing access to datasets requires access to the appropriate infrastructure. Members used and referenced different sources and tools in their posts and conversations, including data, knowledge, and ML model repositories, as well as tools for data access, analysis, annotation, storage, extraction, harvesting, preprocessing, cleaning, sensemaking or visualizing, learning, and searching (see Fig. 5, Table 1). The knowledge of that infrastructure, its components, configurations, and cost is often critical. Members requested and shared infrastructure information to facilitate data sharing and establish and maintain shared community knowledge on big data infrastructure. Some of the core members of the group also built and maintained a concrete infrastructure—a repository—to provide more complete and value-added access to Reddit data than did Reddit's native API. They periodically requested the feedback from the community to make it useful and usable and to increase awareness of it.

[ Fig. 5 about here ]

The analysis of the 100 most frequently referenced domains in the sample identified 21 types or categories of sources. In members' discussions, they referenced social media, software project management, and data-hosting platforms most often (see Table 1). Members shared datasets and information from different government, university, publisher, and digital library sites in addition to Reddit itself and major data-hosting, software management, and data science and machine-learning communities, such as Github and Kaggle. In contrast to the results of a study by Oh et al. (2008) that examined the types of sources used in Yahoo Answers, a general-purpose Q&A website, the differences among the relative frequencies of source categories identified by this study were less sharp. In addition, this study defined a typology with finer granularity for online sources than did Oh and colleagues. That was expected because the study analyzed a more practice-based community, rather than a general-purpose Q&A site. This result also points to the value of examining the sources used by practice-based Q&A and discussion communities.

Typologies like this can be used to annotate and organize data sources for more efficient discovery and access.

[ Table 1 about here ]

The community also discussed professional development and competencies, as well as the skills needed for data work. Some members even shared their mappings of data skills and competencies to data roles when asked for advice on data work-related careers. The professional development- and competency-related themes and actions also included sharing information on books, courses, guides, and tutorials as well as their knowledge of data scraping, big data processing and analysis hardware and software, and machine-learning and data science techniques.

### *5.4. Problems and challenges*

Activity problems or contradictions can be conceptualized as misalignments among the activity components and different activities, including member needs, datasets, and tools (Kaptelinin & Nardi, 2012). The authors categorized problems referenced in the community's discussions into the following general categories: quality, accessibility, ethics, and legality.

#### *5.4.1. Quality*

The community discussed several types of data quality-related problems. Some datasets shared or referenced by members were intrinsically incomplete representations of original phenomena. That made the datasets challenging to use. Members had to engage in data aggregation and enhancement to assemble more complete and useful datasets.

Furthermore, proxy data sharers who obtained data through scraping the original provider's website or calling the provider's API might not have complete copies of data because of the limitations of those methods. In addition, a dataset might not be large enough for some uses. For members to build a high-quality machine-learning model, the training dataset not only had to be representative, but also had to be large enough to successfully apply the machine-learning technique. That is, the dataset had to contain sufficient instances of each relation or concept of the model.

Accuracy is another main dimension of quality. A dataset can be an intrinsically inaccurate representation of an original phenomenon. Data inaccuracy can stem from different sources, including inaccurate data entry, faulty sensors, the use of a biased sample, or purposefully injecting noise into the data. Members often requested information about the provenance of the data. They asked how the data were generated and modified and by whom to predict its accuracy and reliability.

Completeness and accuracy are closely intertwined to shape the quality of data. Bias in data selection can lead to building an incomplete representation of reality, and hence to inaccurate findings and conclusions.

> *You've also limited the data to data from the United States. If you go outside of the United States, you'll notice that other nations have established common-sense gun laws that have had different degrees of effectiveness. . . . This is a terrible misuse of data.* (t104)

Validity and reliability can be considered subcriteria of accuracy. It is difficult to establish data reliability if there is no agreement on how to measure a particular concept or phenomenon. If a dataset does not represent the phenomenon or entity(s) it purports to, then it is not valid. Often members sought confirmation that a dataset was indeed the one they had searched for.

*Is it possible to validate that the download includes the plain XYZ dataset?* (t388)

Indeed, ambiguity was a frequently referenced problem. Members had to make sense of the data first before they could use it. If the data were not accompanied by their schema and other metadata, it could lead to an ambiguity problem. Furthermore, inconsistent use of metadata, data standards, and classification systems could lead to entity misdetermination and accessibility problems, especially if the data were distributed.

*It's a pain to keep an up-to-date processor of their data. They continue to change their standards and conventions without warning.* (t232)

The currency of data is important in many uses. Not many providers, however, would share real-time or current data. The preparation of a dataset for public release, similar to that of software code, can be costly and may take time.

*I wouldn't necessarily anticipate data this quickly if data gathering began in February 2018. It takes a long time to collect data and much longer to prepare the data for dissemination. (t287)*

The understandability of data can be affected by its complexity. Similar to text readability levels, data values could be encoded following a subject-specific coding schema or data transformation technique. The user might need to know that schema and vocabulary to understand or use the data.

*Medical information is also coded. The coding schemas are as complicated as data science itself. Furthermore, understanding any meaning necessitates clinical expertise.* (t609)

The level of data complexity is usually determined by the complexity of the task for which it is created. Members often sought data of a specific complexity for tasks they had to complete, such as school assignments.

*I need to develop an XYZ model for a school assignment, and I need to locate a publicly available dataset. It doesn't matter what it is as long as it is complex enough (approximately nine distinct "classes").* (t653)

In other instances, members commented on the intrinsic complexity of data and what they did to reduce it.

*I also modified the numbers' scale because they were previously shown as whole numbers, which introduced unnecessary complexity when represented as percentages.* (t249)

Thus, as with data quality in general, complexity can be evaluated both intrinsically (e.g., based on the number of variables in its schema and the number of classes in each variable) and relationally (i.e., whether it is more complex than needed for a specific task). Complex problems require complex datasets. Complexity could be seen and discussed as a cost (e.g., making data difficult to understand) and as a value (e.g., enabling the study of complex problems). It could also be seen as a quality problem (e.g., using an unnecessarily complex representation or format for a particular variable); alternatively, it could be seen as a characteristic of higher quality (e.g., using values with high accuracy or ones encoded, preprocessed, or manipulated according to the data representation schemas and practices of a particular research community).

### *5.4.2. Accessibility*

Accessibility was the second most frequently referenced problem after quality-related problems. Members often needed datasets that were not accessible on the open web. They had to scrape and assemble datasets from secondary publications, such as webpages. In other instances, data were available through limited APIs, and members had to work around the restrictions of an API to assemble a more complete dataset. That required members to have considerable coding skills. In

addition, members might not be able to access to data if they did not have the appropriate software to handle the format or scale of the data.

Using nonstandard or proprietary data formats can make data less accessible and interoperable. The community encouraged the use of commonly used data formats when members shared their datasets. Interoperable, nonproprietary data formats are usually simpler and more efficient with regard to storage space. The latter characteristic was particularly important for sharing large datasets.

In addition to data providers explicitly limiting access to their data (e.g., limiting the number of API calls), a sharer might not have access to a suitable infrastructure to provide continuous and effective access to data. Storing, processing, and sharing big datasets could test the limits of members' personal data infrastructure. A large part of the community's deliberation was about finding solutions to data infrastructure challenges.

Finally, members often requested the community's help to obtain data stored behind paywalls, which could involve service account sharing.

> *I have a basic IEEE account. It is not sufficient to just download a dataset. The dataset I'm looking for can be obtained at [URL]. Please help me in gaining access to the dataset.* (t722)

*5.4.3. Ethics*

Data sharing and aggregation can enable intended or unintended unethical uses of data, such as the deanonymization of a person's identity, the development of machine-learning algorithms for surveillance, or, alternatively, the defeating of security and privacy preservation algorithms. The community was aware of those risks and discussed them frequently.

> *I know it's entertaining, but is it really ethical to distribute this data? By making datasets with captcha available public, you're offering ammunition to individuals who wish to develop captcha-solving bots.* (t292)

A tradeoff exists between accessibility and privacy. The lack of access can also be an ethical issue because it can lead to a digital divide or a less informed citizenry, or both.

> *The sole reason for suppression rules is that they are probably required by some mandates. . . . Obtaining county level data necessitates the analyst group to have as a Health Care Authority status. Individual analysts, I believe, are experiencing a dataset "Dark Age."* (t981)

Another tradeoff is between privacy and data quality. Some members of the community expressed concerns about the quality of modified data. Users modify data to make it amenable to a particular analysis or to comply with specific policies, regulations, and laws. One of the motivations for modifying data is to preserve people's privacy. That is usually accomplished by injecting noise or inaccuracy into an original dataset. However, injecting noise into data to protect individuals' privacy can undermine the statistical validity of the data or render the data unusable for existing purposes, or both. This practice can be particularly impactful for national survey-based datasets that are used in both research and policy design.

Members often reverse-engineered datasets by scraping them from websites. In those cases, they might not own the data or have explicit permission to use or share it. Members sought and received advice from the community on harvesting, providing access, and using third-party data. As expected, sharers were interested in the legality of extracting and sharing third-party data, whereas seekers were interested in the legality of using the data. The community also discussed

and offered advice on how to work around and protect oneself from potential legal problems raised by unauthorized data sharing and the use of third-party data.

> *If you're worried about it being published, simply don't include your name or other personal information. It is unlikely that you'd be singled out for aggregating a dump that's already out there.* (t26)

Unauthorized data use, aggregation, and sharing can lead to both ethical and legal problems (Mason, 1986; California Consumer Privacy Act [CCPA] (https://datasetsearch.research.google.com/), General Data Protection Regulation ([GDPR] https://gdpr.eu/). Hence, dataset descriptions should specify the provenance of the data, including its source, owner, and method of generation. That is, data creators can reduce challenges and problems with data use and sharing downstream by making the process of dataset creation more transparent through documentation. Likewise, because many of those datasets are used in teaching and research, librarians should provide their user groups with guides and training on the ethical and legal aspects of third-party data use and sharing.

*5.5 Implications for the design of dataset metadata: Data Q&A metadata (DQAM) model*

Metadata is critical to successful dataset sharing and reuse. This section reports on the study's metadata related findings and their implications. The implications are presented as a metadata model to make them more actionable and usable for the design of dataset metadata schemas and vocabularies.

The examination of members' actions, challenges, and types of information and data requested, shared, or discussed in the Q&A and dissemination activities suggests 16 metadata elements that should be included in a descriptive metadata model for datasets (see Table 2). The paper refers to this set of elements hereafter as the data Q&A metadata (DQAM) model. The Quality element has the most detailed model. The DQAM model also includes Provenance-related elements that, when combined with Quality elements, can support quality evaluation and credit allocation. The Privacy Risk element encodes information on different levels of privacy risks (e.g., non-human-subject vs. human subject data; anonymized vs. pseudo anonymized data). Data quality and privacy are critical ethical issues for information systems (Mason, 1986), and the findings of this study corroborate this fact. It is important that data repositories and search engines such as Google Data Search enable users to search for and select datasets that meet their data quality and privacy preferences or datasets that are compliant with state and national privacy laws and regulations (e.g., CCPA, GDPR). This is especially relevant when users search for and share datasets on the uncurated web, where users might not have the traditional guidance and advice on ethical issues of data management that are available at research universities and in large research laboratories (e.g., institutional review boards). The management of technical aspects is just one side of data curation. Another side is the management of legal and ethical issues. Although license or property information has been a part of standard metadata schemas and vocabularies for a very long time, privacy and quality have not. In the era of disinformation and misinformation (Lazer et al., 2018; Stvilia, 2021), privacy and quality facets of dataset description become increasingly crucial for facilitating the ethical use and sharing of datasets. Furthermore, as this analysis shows, users may be interested not only in high-quality datasets, but also in noisy, low-quality datasets so that they can teach or practice data cleaning and wrangling methods.

The Value element of the DQAM model can be used to communicate the importance of a dataset to the community. It can be measured by the frequency of use or community-approved measures of value, such as the number of upvotes. The Related Objects element can include

pointers to objects related to a dataset, such as the project that produced or used the dataset, a data paper that describes the dataset, or other datasets that complement or have been used with the dataset. The Related Objects element can help with both making sense of the dataset and evaluating its quality. The information encoded in the Scale/Size element can be used to assess the infrastructure needs for a dataset. The user may also use that metadata in combination with the Completeness and Complexity subelements to determine the applicability of a dataset for a particular task, such as training a machine-learning algorithm or meeting the requirements of a school assignment.

The authors mapped the DQAM model onto the four facets of the FAIR framework: findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016; see Table 2). The mapping showed that the DQAM model supported all those facets. The Quality element of the DQAM can be used in operationalizing all four dimensions of the FAIR. The Provenance and Related Objects elements can support the discovery and reusability of data. The quality of an information object can be evaluated directly by examining the object itself, or indirectly by examining the process of its creation and manipulation (Stvilia et al., 2007). The content of the Provenance and Related Objects elements of the DQAM model can be utilized in assessing the quality of the processes used to create and modify a particular dataset. For instance, providing information about the annotators of a dataset could help predict the credibility and political biases of the annotations to a dataset (Scheuerman et al., 2021).

[ Table 2 about here ]

*5.6 Limitations and future research*

The study has one limitation. The findings were derived by examining the log of a single data discussion community. A future research related study will collect additional data by interviewing members of the community to expand and triangulate the findings of this study. This study did not have access to private conversations and discussions of the group moderators. Hence, the study findings regarding the community design and maintenance activity are very limited. Future research could interview moderators of the group to develop a more complete understanding of the community design and maintenance activities of the subreddit.

## 6. Conclusion

Open access to user discussions and Q&A on social media platforms enables researchers to examine at scale the data management needs and behaviors of dataset creators, sharers, and users on the Web. This study examined discussions of the r/Datasets community on Reddit using a methodology comprised of activity theory, content analysis, and data visualization. Members discussed how to find, create, obtain, and aggregate datasets. They used 21 types of information and data sources in their activities. Additionally, the study developed a metadata model for datasets and map it onto the four facets of the FAIR framework.

The findings of this study enhance our understanding of the activity structures, data and information sources used, and challenges and problems encountered when users search for, share, and make sense of datasets on the web, outside the traditional information and data ecosystems. Data librarians and curators can use the findings of this study in the design of their data management and reference services. The typology of data sources and the metadata model developed through this study can be used in annotating and categorizing data sources and informing the design of metadata schemas and vocabularies for datasets.

**References**


Adamic, L.A., Zhang, J., Bakshy, E., & Ackerman, M.S. (2008). Knowledge sharing and *Yahoo! Answers*: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 665–674). New York, NY: ACM.

Auxier, B., & Anderson, M. (2021). *Social media use in 2021*. Washington, DC: Pew Research Center. Retrieved January 6, 2022, from https://www.pewresearch.org/internet/2021/04/07/acknowledgments-38/

Bailey, K.D. (1994). *Typologies and taxonomies: An introduction to classification techniques*. Thousand Oaks, CA: Sage.

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13,* 407–424. https://doi.org/10.1108/eb024320

Birnholtz, J.P., & Bietz, M.J. (2003, November). Data at work: Supporting sharing in science and engineering. In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 339–348). New York, NY: ACM.

Borgman, C.L., Wallis, J.C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries, 7,* 17–30.

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy, 33,* 599–616.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15,* 662–679. https://doi.org/10.1080/1369118X.2012.678878

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*, 249–254.

Chen, S., & Chen, B. (2020). Practices, challenges, and prospects of big data curation: A case study in geoscience. *International Journal of Data Curation, 14,* 275–291.

Clement, J. (2021). Reddit users: Unique monthly visits 2021. *Statista.* Retrieved January 6, 2022, from https://www.statista.com/statistics/443332/reddit-monthly-visitors/

Coletta, R., Castanier, E., Valduriez, P., Frisch, C., Ngo, D., & Bellahsene, Z. (2012, May). Public data integration with websmatch. In *Proceedings of the First International Workshop on Open Data* (pp. 5–12). New York, NY: ACM.

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2007, January). SuggestBot: Using intelligent task routing to help people find work in Wikipedia. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (pp. 32–41). New York, NY: ACM.

Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368,* 4023–4038.

Cunningham, S.J., & Hinze, A. (2014). Social, religious information behavior: An analysis of Yahoo! Answers queries about belief. *Advances in the Study of Information and Religion, 4,* 1–26. https://doi.org/10.21038/asir.2014.0002

Dash, S., Shakyawar, S.K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data, 6,* 1–25.

Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A., & De Laat, C. (2012, December). Addressing big data challenges for scientific data infrastructure. In *4th IEEE*

*International Conference on Cloud Computing Technology and Science Proceedings* (pp. 614–617). Piscataway, NJ: IEEE.

Ferguson, A.R., Nielson, J.L., Cragin, M.H., Bandrowski, A.E., & Martone, M.E. (2014). Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience, 17,* 1442–1447.

Fu, H. (2019). *Peer production of knowledge in online social Q&A communities at startup stage*. Retrieved January 7, 2022, from http://purl.flvc.org/fsu/fd/2019_Spring_Fu_fsu_0071E_15032

Fu, H., & Stvilia, B. (2016, June). Knowledge curation discussions and activity dynamics in a short lived social Q&A community. In *2016 IEEE/ACM Joint Conference on Digital Libraries* (*JCDL*) (pp. 203–204). Piscataway, NJ: IEEE.

Fu, H., Wu, S., & Oh, S. (2015). Evaluating answer quality across knowledge domains: Using textual and non-textual features in social Q&A. *Proceedings of the Association for Information Science and Technology, 52*, 1–5.

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75,* 861–871.

Gutmann, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. *Data Science Journal, 3,* 209–221.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3,* 134–140. https://doi.org/10.2218/ijdc.v3i1.48

Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. (2012). Prioritization of data quality dimensions and skills requirements in genome annotation work. *Journal of the American Society for Information Science and Technology, 63,* 195–207. https://doi.org/10.1002/asi.21652

Juran, J. (1992). *Juran on quality by design*. New York: The Free Press.

Kaptelinin, V., & Nardi, B. (2012). Activity theory in HCI: Fundamentals and reflections. *Synthesis Lectures on Human-Centered Informatics, 5*, 1–105. https://doi.org/10.2200/S00413ED1V01Y201203HCI013

Laskowski, C. (2021). Structuring better services for unstructured data: Academic libraries are key to an ethical research data future with big data. *The Journal of Academic Librarianship, 47*, Article 102335.

Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., & Zittrain, J.L. (2018). The science of fake news. *Science, 359,* 1094–1096.

Lee, D.J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PloS One, 12*, Article e0173987.

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., & Hartmann, B. (2011, May). Design lessons from the fastest Q&A site in the west. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2857–2866). New York, NY: ACM.

Mason, R.O. (1986). Four ethical issues of the information age. *MIS Quarterly, 10,* 5–12.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. Washington, DC: National Academies Press.

Ng, A. (2021, July 29). AI doesn't have to be too complicated or expensive for your business. *Harvard Business Review*. Retrieved January 7, 2022, from https://hbr.org/2021/07/ai-doesnt-have-to-be-too-complicated-or-expensive-for-your-business

Oh, S., Oh, J.S., & Shah, C. (2008). The use of information sources by internet users in answering questions. *Proceedings of the American Society for Information Science and Technology, 45*(1), 1–13.

Pal, A., Harper, F.M., & Konstan, J.A. (2012). Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Transactions on Information Systems, 30*(2), 1–28. https://doi.org/10.1145/2180868.2180872

Pirolli, P. (2009, April). An elementary social information foraging model. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 605–614). New York, NY: ACM.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 248–256). Stroudsburg, PA: Association for Computational Linguistics.

Scheuerman, M.K., Hanna, A., & Denton, E. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human–Computer Interaction*, *5*(CSCW2), 1–37.

Stvilia, B. (2021). An integrated framework for online news quality assurance. *First Monday, 26*(7). https://doi.org/10.5210/fm.v26i7.11062

Stvilia, B., Hinnant, C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2017). Toward collaborator selection and determination of data ownership and publication authorship in research collaborations. *Library & Information Science Research*, *39*(2), 85-97.

Stvilia, B., Hinnant, C., Wu, S., Worrall, A., Lee, D. J., Burnett, K., Burnett, G., Kazmer, M. M., & Marty, P. F. (2015). Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology*, *66*(2), 246-263.

Stvilia, B., Hinnant, C., Schindler, K., Worrall, A., Burnett, G., Burnett, K., Kazmer, M., & Marty, P. (2011). Composition of scientific teams and publication productivity at a national science lab. *Journal of the American Society for Information Science and Technology*, *62*(2), 270-283.

Stvilia, B., Gasser, L., Twidale M., B., Smith L. C. (2007). A framework for information quality Assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720-1733.

Stvilia, B., Twidale, M., Smith, L. C., Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983-1001.

Swarup, S., Braverman, V., Arora, R., Caragea, D., Cragin, M., Dy, J., Honavar, V., Huang, H., Locicero, R., Singh, L., & Yang, C. (2018, June). Challenges and opportunities in big data research: Outcomes from the second annual joint PI meeting of the NSF BIGDATA research program and the NSF big data regional innovation hubs and spokes programs 2018. In *NSF Workshop Reports* (pp. 1–9). Alexandria, VA: National Science Foundation. Retrieved January 7, 2022, from https://par.nsf.gov/biblio/10113364-challenges-opportunities-big-data-research-outcomes-from-second-annual-joint-pi-meeting-nsf-bigdata-research-program-nsf-big-data-regional-innovation-hubs-spokes-programs

Tausczik, Y.R., Kittur, A., & Kraut, R.E. (2014). Collaborative problem solving: A study of math overflow. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 355–367). New York, NY: ACM.

Taylor, R.S. (2015). Question-negotiation and information seeking in libraries. *College & Research Libraries, 76,* 251–267.

Tenopir, C., Sandusky, R.J., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research, 36,* 84–90.

Vicente-Saez, R., Gustafsson, R., & Martinez-Fuentes, C. (2021). Opening up science for a sustainable world: An expansive normative structure of open science in the digital era. *Science and Public Policy*, *48*, 799-813.

Wiggins, A., & Crowston, K. (2015). Surveying the citizen science landscape. *First Monday, 20*. Retrieved January 7, 2022, from https://firstmonday.org/article/view/5520/4194

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 1–9.

Wilson, T.D. (1997). Information behaviour: An interdisciplinary perspective. *Information Processing & Management, 33,* 551–572.

Wilson, T.D. (2006). A re-examination of information seeking behaviour in the context of activity theory. *Information Research: An International Electronic Journal, 11*(4), Article n4.

Wu, S., & Worrall, A. (2019). Supporting successful data sharing practices in earthquake engineering. *Library Hi Tech, 7,* 764–780. https://doi.org/10.1108/LHT-03-2019-0058

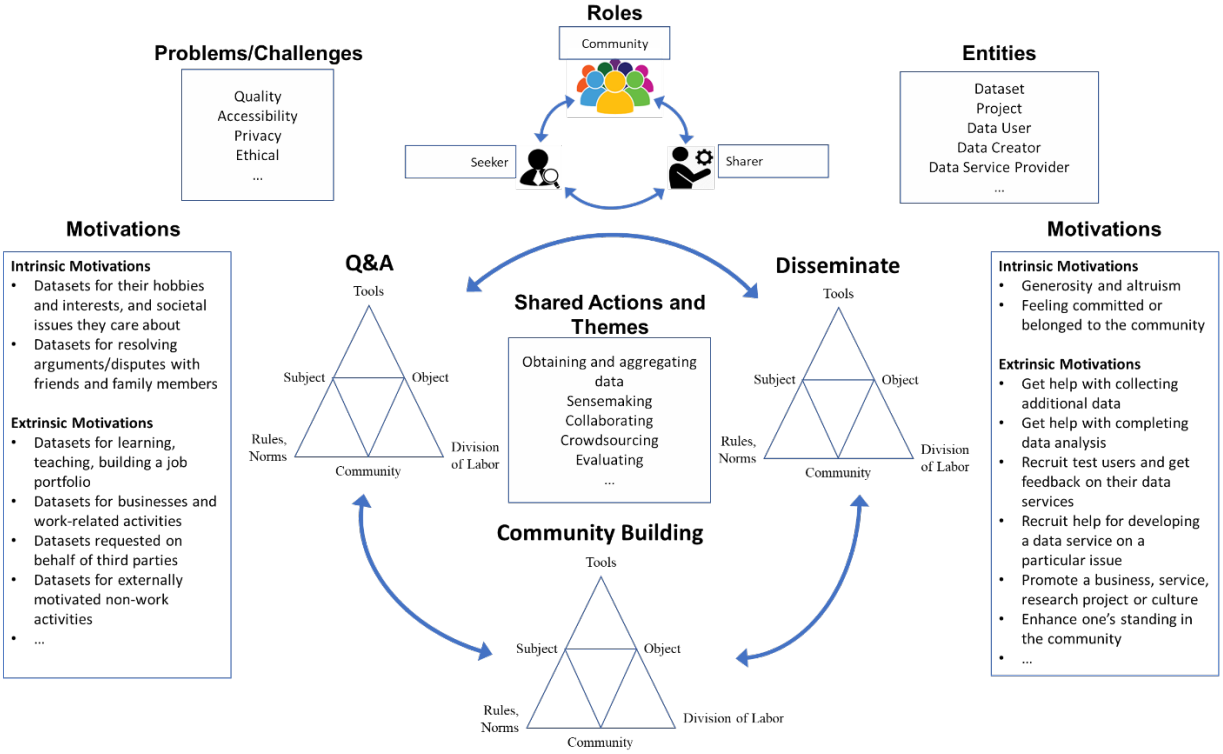

Fig. 1. Activity structure of the community.

| | |
|---|---|
| Q&A | looking, dataset, request, state, county, anyone, search, information, thanks, country, model, population, census, weather, national, source, hello, income, location, company, survey, scrape, election, everyone, college, visualize, death, specific, price, health, school, similar, sale, train, finding, average, result, found, variable, vote, advance, city, historical, global, public, zillow, wondering, california, local, analyze, database, student, number, contains, trying, region, item, temperature, image, statistic, police, economic, university, space, year, government, cancer, legal, education, annual, property, proportion, disaster, effect, appreciate, amazon, place, covid, service, world, least, urban, climate, question, household, somewhere, detection, indicator, cite, recipe, shooting, production, status, bureau, disease, employee, precipitation, searching, energy, career |

Fig 2. The term profile of the question-and-answer activity.

| | |
|---|---|
| Dissemination | comment, reddit, thanks, dataset, submission, tweet, search, torrent, month, file, available, share, million, bigquery, download, subreddit, script, awesome, count, server, score, column, using, compressed, object, interested, query, language, complete, link, meme, sentiment, please, format, total, upload, monthly, limit, issue, uncompressed, update, example, amazon, repository, elasticsearch, twitter, removed, space, false, improve, updated, entire, archive, dolthub, stream, code, synthetic, original, collected, index, missing, contains, hour, magnet, crawler, period, article, dump, metadata, api, structure, title, endpoint, jeopardy, review, deleted, mysql, collecting, variable, entry, delimited, filtering, corpus, clone, handle, unique, snapshot, approximately, billion, recipe, filter, captured, parent, mirror, double, project, edited, pushshift, respect, paper |

Fig. 3. The term profile of the dissemination activity.

| | |
|---|---|
| Community Building | driven, scientist, people, financial, discussion, housing, chat, survey, happiness, synthetic, pitch, fun, political, favorite, white, synesthesia, reddit, statistical, community, transformation, economic, mortgage, biased, disease, similarity, submission, policy, subtitle, playlist, antibody, vector, subprime, matrix, effect, group, color, default, crisis, expectancy, macro, complain, titanic, interest, conference, monthly, government, increasing, existed, programmer, statistically, pacer, climate, loan, population, platform, privacy, distance, trulia, inequality, regional, slavery, random, politics, racist, encoding, write, measurable, ignored, empirical, affiliation, lecture, prove, lending, disaster, impact, february, racialized, transition, colour, destroy, labor, chain, recurrence, thread, vaccination, bureau, denial, associate, quartet, antigen, plague, sticky, outrageous, parody, repayment, decline, foia, research, play, industry |

Fig 4. The term profile of the community-building activity.

| | |
|---|---|
| Tools, infrastructure | python, tableau, comment, file, subreddit, memory, analyst, server, selenium, order, total, select, learn, title, word, distinct, software, byte, input, phrase, schema, dump, tensorflow, corpus, textual, ggplot, powerbi, count, plotly, bittorent, cloud, favorite, matplotlib, group, drive, solr, field, pytorch, lake, plot, backup, modelling, core, static, dropbox, seaborn, compression, fitness, written, corporate, unicode, mine, encountered, sas, dedicated, commenters, elasticsearch, permalink, stackexchange, javascript, warehousing, animation, text, hadoop, titan, samsung, incels, jmp, beautifulsoup, moment, harvest, apache, toshiba, enterprise, powershell, sharding, r_language, database, bigquery, amazon, sqlite, airflow, pandas, storage, github, ec2, spotfile, autocad, matlab, reddit, ubuntu, bokeh, census, metadata, algoria, kaggle, csv, requests, pyplot, scientist |

Fig 5. The term profile of tools.

Table 1. Categories of information and data resources referenced in the sample.

| No. | Categories of sources | Frequency | % |
|---|---|---|---|
| 1 | Social media platforms (e.g., Reddit) | 647 | 26.9 |
| 2 | Software project management platforms (e.g., Github) | 431 | 17.9 |
| 3 | Data-hosting and data-sharing platforms (e.g., Google Cloud) | 325 | 13.5 |
| 4 | AI communities and companies (e.g., Kaggle) | 219 | 9.1 |
| 5 | Government agencies (e.g., Census.gov) | 182 | 7.6 |
| 6 | Digital libraries (e.g., Arxiv.org) | 153 | 6.4 |
| 7 | Community data repositories and data companies (e.g., Pushshift) | 91 | 3.8 |
| 8 | Universities and research laboratories (e.g., uci.edu) | 85 | 3.5 |
| 9 | Software companies and communities (e.g., Wolframalpha) | 66 | 2.7 |
| 10 | Encyclopedias (e.g., Wikipedia) | 62 | 2.6 |
| 11 | Online publishing or blogging companies (e.g., Medium) | 53 | 2.2 |
| 12 | International organizations, consortia, and standardization bodies (e.g., W3C) | 14 | 0.6 |
| 13 | Newspapers (e.g., *Washington Post*) | 14 | 0.6 |
| 14 | Banks (e.g., Federal Reserve Bank of St. Louis) | 12 | 0.5 |
| 15 | Publishers (e.g., Nature) | 12 | 0.5 |
| 16 | Research information management systems (e.g., Researchgate) | 10 | 0.4 |
| 17 | Ontologies (e.g., Wikidata) | 9 | 0.4 |
| 18 | Real estate marketplace companies (e.g., Zillow) | 9 | 0.4 |
| 19 | Opinion poll analysis companies (e.g., Fivethirtyeight) | 5 | 0.2 |
| 20 | Learning companies (e.g., Datacamp) | 4 | 0.2 |
| 21 | Sports leagues (e.g., NBA) | 4 | 0.2 |

**Table 2.** Data Q&A Metadata (DQAM) model. Notes. FAIR = **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability. Letters in the second column indicate which FAIR dimension a particular DQAM model element contributes to.

| **DQAM Model Elements** | **Mapping to FAIR** |
|---|---|
| 1. Citation | F |
| 2. Cost (total cost of management) | R |
| 3. Description | F,R |
| 4. Format | F,I |
| 5. Identifier | F |
| 6. License/Rights | F,A |
| 7. Privacy Risk (i.e., level of privacy risk) | A,R |
| 8. Provenance | F,R |
| 8.1 Provenance: Owner | |
| 8.2 Provenance: Creator | |
| 8.3 Provenance: Sharer | |
| 8.4 Provenance: Modifier | |
| 8.5 Provenance: Source | |
| 8.6 Provenance: Creation Method | |
| 8.7 Provenance: Date | |
| 8.8 Provenance: Version; Audit Trail | |
| 9. Quality | F,A,I,R |
| 9.1 Quality: Accessibility—interoperability of the data format used; data sharing restrictions | |
| 9.2 Quality: Accuracy—invalid measurements; inaccurate measurements | |
| 9.3 Quality: Completeness—number of variables; number of instances | |
| 9.4 Quality: Complexity—number of variables; number of classes for each variable; "data readability level" (level of subject expertise needed to understand data) | |

| | |
|---|---|
| 9.5 Quality: Consistency—consistency of data schema; consistency of value encoding | |
| 9.6 Quality: Currency | |
| 10. Related Objects/Context | F,R |
| 10.1 Related Code (software used to generate, process, or analyze the data) | |
| 10.2 Related Dataset | |
| 10.3 Related Paper | |
| 10.4 Related Project Description | |
| 10.5 Related Project Title | |
| 11. Scale/Size | F,R |
| 12. Schema/Variables | F,I,R |
| 13. Subject | F |
| 13.1 Subject: Activity | |
| 13.2 Subject: Event | |
| 13.3 Subject: Location | |
| 13.4 Subject: Population | |
| 13.4 Subject: Subject | |
| 13.5 Subject: Time Period | |
| 14. Title | F |
| 15. Type (e.g., numerical, categorical, time series) | F,R |
| 16. Value | A |