The published version is available:
Haiyan Jia, Larrisa I. Miller, Jessica Hicks, Ethan Moscot, Alissa Landberg, Jeff Heflin & Brian D. Davison (2022) Truth in a sea of data: adoption and use of data search tools among researchers and journalists, Information, Communication & Society, DOI: 10.1080/1369118X.2022.2147398

Truth in a Sea of Data: Adoption and Use of Data Search Tools among Researchers and Journalists

Haiyan Jia[a]*, Larrisa Miller[b], Jessica Hicks[c], Ethan Moscot[d], Alissa Landberg[d], Jeff Heflin[e], and Brian Davison[e]

[a]Department of Journalism and Communication, Lehigh University, Bethlehem, PA, USA; [b]Department of Communication, University of Massachusetts Amherst, Amherst, MA, USA; [c]Independent research, New York, NY, USA; [d]Cognitive Science Program, Lehigh University, Bethlehem, PA, USA; [e]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA

*Email: haiyan.jia@lehigh.edu

Haiyan Jia is an assistant professor in the Department of Journalism and Communication at Lehigh University. Her research interest primarily focuses on the social and psychological effects of communication technology ranging from the Web to mobile apps to smart objects. She also researches how technology advances data journalism and shapes privacy perceptions and behaviors.

Larrisa (Larri) Miller is a PhD student in the Department of Communication at the University of Massachusetts Amherst. Her research primarily focuses on the digital information ecosystem. She is interested in misinformation flow, political communication, computational social science, and feminist theory.

Jessica Hicks is an independent researcher based in New York City. She graduated from Lehigh University with a Bachelor's Degree in Journalism, Sociology, and Anthropology, and researched data accessibility and its implications for journalism and society at large. Her current work focuses on mental health, well-being, and behavior change in the digital health technology industry.

Ethan Moscot is an undergraduate student in the Cognitive Science Program at Lehigh University. His research interests primarily lie at the intersection of data and policy. His current research explores the technical, political, and legal factors related to information privacy.

Alissa Landberg is a senior at Lehigh University majoring in both Cognitive Science and Computer Science. Her research interests focus on the interdisciplinary relationship between behavioral and technical studies. Within these fields, she focuses on collective privacy

dimensions within smart home environments and technical improvements and user perspectives to enhance dataset search.

Jeff Heflin is an associate professor in the Department of Computer Science and Engineering at Lehigh University. His specific research interests include machine learning for dataset search and augmentation, establishing semantic interoperability between heterogeneous information systems, exploration and analysis of complex data, scalable ontology reasoning, and developing formal theories of distributed ontology systems.

Brian D. Davison is a professor in the Department of Computer Science and Engineering at Lehigh University and provides leadership to the university's interdisciplinary data science programs. He also serves as associate director of Lehigh's Institute for Data, Intelligent Systems and Computation (I-DISC). His research includes search, mining, recommendation and classification problems in text, on datasets, on the Web and in social networks.

Truth in a Sea of Data: Adoption and Use of Data Search Tools among Researchers and Journalists

The increasing availability of data search tools brings opportunities for non-expert users. Among these users, interdisciplinary researchers and data journalists represent a growing population whose work can lead to societal benefit. Through in-depth interviews, we examine what strategies and approaches researchers and journalists adopt to search online data, how they apply current technology to facilitate dataset search, and the barriers and difficulties that they encounter in their work with data. Our findings reveal that with technological limitations in the aspects of searchability, interactivity and usability, dataset search for non-experts remains a challenge. We have found that little attention has been paid to non-experts' emerging data need, leading to significant constraints to the design and development of technological tools for supporting non-expert users. Our findings underline the critical impact of the design, development and deployment of technological tools to enable the meaningful use of today's increasingly available data toward a civil society.

Keywords: non-experts, dataset search, searchability, interactivity, usability

1. Introduction

The number of websites and online services hosting and indexing large datasets has increased over the past decade, which seemingly make data free and accessible for anyone willing to look for them. We now see a growing population of users utilizing these resources to access publicly available data. These emerging 'casual' users tend to find Web-based information seeking much more laborious and frustrating than expert users (Hölscher & Strube, 2000). While previous research has shown that lack of knowledge structures and strategies negatively affect Web search behavior of novice users, it is unclear if we will find similar effects on non-experts' adoption and use of dataset search tools. In the context of dataset search, individuals who lack Web and search experiences, data skills, or domain-specific knowledge, could all be considered

as non-experts. In this study, we have identified two groups who represent non-expert users: researchers working in interdisciplinary areas and data journalists. Both groups use online data search tools in their work and lack relevant skills and expertise. We have investigated how they search data online, what obstacles they encounter, and how these obstacles affect their use and adoption of these tools as well as their ability to reveal insights from data.

1.1 Data Accessibility as a Misconception

The capacity to capture massive amounts and new types of data has transformed scientific research in many disciplines. Data has been increasingly adopted by scholars in fields that do not typically utilize large-scale datasets. Researchers who work on these cross-domain projects may lack in-depth knowledge about the new disciplinary field or data science. Data search tools accessible to these researchers are very complex, are not optimized for Open Data, and do not address the need for flexibility to perform search tasks across multiple datasets, systems, or domains (Braunschweig, Eberius, Thiele, & Lehner, 2012). This creates great challenges to accessing relevant and reliable data.

Similarly, the emergence of data journalism as a field is a recent development (Gray, Chambers, & Bounegru, 2012). Researchers have claimed that journalistic practices have taken 'a quantitative turn' as data have become more important and prevalent in professional journalism (Coddington, 2015). Reviews of award-winning data journalistic projects in the past few years (e.g., Loosen, Reimer, & De Silva-Schmidt, 2020; Young, Hermida, & Fulda, 2018) have revealed that, while the majority of projects rely on readily accessible datasets primarily released from official institutions, journalists are increasingly looking for unofficial data sources for their stories. Studies have found that it is challenging for journalists to locate and identify

data usable and useful to their work (Noy & Brickley, 2017), and the idea of data already being widely obtainable is a 'misconception' (Kitchin, 2014).

1.2 Challenges for Non-Experts

Data-driven technology and tools have become more readily available, ranging from domain-specific data repositories (e.g., Ag Data Commons) and data indices (e.g., Data Is Plural), to data centers of governmental agencies (e.g., data.gov) and commercial data hubs (e.g., Google Dataset Search). However, they are not easy to use, especially for non-expert users.

Non-experts consist of a variety of individuals and groups, as expertise is a multi-faceted concept with *technical*, *informational* and *experiential* aspects (e.g., see Duggan & Payne, 2008; Jenkins, Corritore, & Wiedenbeck, 2003; White, Dumais, & Teevan, 2009). Technical expertise spans from general skills such as web or media competence and literacy (Hölscher & Strube, 2000) to specialized skills, in this case, search expertise (Umemoto, Yamamoto, & Tanaka, 2016) and data analytic and data assessment skills (Espinosa, García, Zorrilla, Zubcoff, & Mazón, 2013). Information expertise refers to domain-specific background knowledge (White et al., 2009), including topic familiarity (Kelly & Cool, 2002), domain vocabulary (Heflin, Davison, & Jia, 2021; Nguyen, Rybinski, Karimi, & Xing, 2022), and domain-specific search expertise (Mao, Liu, Kando, Zhang, & Ma, 2018). Lastly, related experience, especially domain-specific search experience, significantly predicts search effectiveness; such experiences help build mental models on which individuals rely to form effective searches (Slone, 2002).

Research has indicated that non-experts typically go through multi-step processes to identify search strategies, sources, and domains appropriate for their work

(Gray et al., 2012). Non-experts also have more difficulty adopting and applying advanced data techniques, formulating queries to meet search needs, and building reliable expectations of a novel system (Zahidi, Lim, & Woods, 2014; Nguyen et al., 2022). The lack of functional, usable and interactive dataset search tools limits the speed and scope of scientific discovery, and journalists' role in facilitating and gate-keeping information toward a civil society (Baack, 2018).

Open data means nothing if it is inaccessible. Although a 'data rush' has incentivized scholars and journalists to work with data, most of the time, '[t]he material is not indexed in any meaningful way', nor can non-experts verify data quickly and easily (Mahrt & Scharkow, 2013). This makes it difficult to ensure reliability, accuracy and clarity in a timely fashion in scientific or journalistic work (Cushion, Lewis, & Callaghan, 2017).

1.3 Limitations of Existing Technology

Research that reviews the technological tools at the disposal of non-experts deems that they 'fall very short of expectation' (Bonaque et al., 2016). For instance, governmental agencies do not see journalists as the end users of their data. Therefore, their data releases, 'even in the best cases, are uneven, and slow, and do not meet the needs of journalists' (Stoneman, 2015). On the other hand, unofficial datasets tend to be equally 'unusable, due to the fact that the design of such tools rarely consider its applicability (to these emerging fields), or the skills, needs, and preferences (of these non-expert users)' (Loosen et al., 2020).

Researchers and data journalists not only face technical barriers, but also lack domain expertise. The latter prohibits them from knowing credible data sources in these fields, accessing domain-specific search tools, and constructing effective search queries. In some domains (e.g., social sciences), large-scale data repositories are almost non-

existent, and collaboration on data collection and data-sharing is also absent between different disciplines (Schroeder, 2020).

1.4 Designing for Non-Expert Data Search

An effective solution requires a user-centric approach that focuses on the preferences and needs of non-expert users in domain-agnostic data search, leading to innovative interface designs and affordances that can fundamentally enhance the usability and ease of use of data search tools. Research on data search shows that common document retrieval techniques used in web search are not optimized for Open Data and do not support non-expert use of the system (Braunschweig et al., 2012).

More broadly, we refer to the rich body of work on human-computer interaction (HCI) and computer-mediated communication (CMC), which has produced vast amounts of insights for user-centric designs. HCI and CMC research has shown that technological affordances such as interactivity and navigability may facilitate information processing. For instance, highly *interactive* user interfaces empower non-expert users by enabling two-way communication, user control, and less time to find the needed information (McMillan & Hwang, 2002). In journalism research, high interactivity is associated with perceived credibility, as reciprocal message exchanges lead to greater gratification and positive attitudes toward the content (Jahng & Littau, 2016). Similar findings also appear in research related to data use. Interactivity is found to facilitate content filtering, getting notifications, and navigation through external sources (Zelenkauskaite & Simões, 2014). Similarly, navigability is found to have a positive effect on memory of site content and content attitudes toward the site (Sundar, Jia, Waddell, & Huang, 2015).

An extensive review of the literature reveals new opportunities for developing effective dataset search tools for non-expert users. For instance, we consistently find

that state-of-the-art data search engines especially lack CMC tools that enable interpersonal interaction or user choice. The existing literature has shown that CMC affordances allow individuals to 'have the capacity to take a more active role in information consumption' (Tremayne & Dunwoody, 2001). We propose to take a user-centric approach toward understanding non-expert users' adoption and use of data search tools, which is critical for designing technological tools that empower this growing user population.

In this paper, we aim to answer the following research questions:

RQ1: How do non-expert users search, assess, and gather data online?

RQ2: What challenges and obstacles do non-experts encounter in dataset searches? How would these challenges, in turn, affect their work and data-driven discoveries?

RQ3: What factors can positively affect non-experts' adoption and use of dataset search tools?

In order to address these research questions, we conducted in-depth interviews with researchers and data journalists, asking them about their past experiences, present objectives, and visions for the future. Our study mainly focuses on structured quantitative data, while recognizing the diversity of data types and formats as well as the implications of such diversity for data search practices. Research methods and findings will be discussed in the sections below.

2. Method

2.1. Participants

Semi-structured, in-depth interviews were conducted to address the research questions above. We identified researchers and data journalists as the target interviewees, as they

represent individuals who frequently utilize data for professional purposes, with limited background and training in related domains.

The recruitment process started with compiling a list of researchers who are involved in interdisciplinary projects in several institutions and a list of data journalists in nationally renowned news organizations and local newsrooms. We emphasized the interdisciplinarity and cross-domain work of the potential participants for their non-expert perspectives. For instance, a computer scientist or a librarian would be a non-expert user if their work required finding datasets in an unfamiliar, niche domain. In total, we contacted over 250 researchers and journalists with information about the objectives and the procedure of the study. We completed interview sessions with 24 participants, including 20 researchers and 4 journalists, who all self-identified as conducting interdisciplinary or cross-domain work (see Appendix A for details including participant numbers). 12 were conducted in person, and the other 12 were conducted remotely via Skype or Zoom. The interviews lasted 30 to 60 minutes, and were audio recorded and then transcribed manually or via Otter.ai, a speech-to-text transcribing service.

2.2. Procedure and Instrument

Participants were asked to provide their consent before completing the survey questionnaire. Questions were compiled through extensive literature review (e.g., Al-Maskari & Sanderson, 2011; Borges-Rey, 2016; Fink & Anderson, 2015; Loosen et al., 2020; Suhr, Dungl, & Stocker, 2020; Wongsuphasawat, Liu, & Heer, 2019, etc.) and pretested with domain experts. Topics covered by the interview questions include the nature of participants' work and their approach to searching online datasets (e.g., 'Does your work involve data, and if so, how?' 'Where do you usually go to find the data?' and 'What does your workflow look like when you are collecting data?'), experiences

with currently available data websites and dataset search tools (e.g., 'How long does it take you to find relevant data sets?' 'How long does it take you to complete a project?' and 'How do you evaluate the reliability of the data?'), and their thoughts on what could improve the interface design and functionality of data set search engines and tools (e.g., 'Can you describe a time you were frustrated by a search engine?' 'Are available tools and search engines meeting your data-collection needs?' and 'What kinds of tools are missing?' etc.).

2.3. Analysis

Anonymized interview transcripts were first analyzed using a bottom-up, qualitative approach (Charmaz, 2006; Glaser & Strauss, 1967; Lofland, Snow, Anderson, & Lofland, 1971). A portion of the transcripts were iteratively read and reread to identify patterns and themes in participants' responses. This identification process involved both independent analysis and group discussions. After identifying salient themes, we aligned the themes with the primary research questions by examining participants' strategies for finding and working with data, problems and issues, and features and processes that facilitate data practices.

A codebook was generated based on our research questions and the emerging patterns. It contained seven major themes, each consisting of a number of sub-themes with qualitative categories, e.g., Background Information (domain, expertise); Work flow (process, data acquisition, time sensitivity, collaboration); Source (source type, reliability, frequently used sources); Search (search process, frequently used queries, matching results); Interaction (search refinement, parameters, data visualization, communication); Usability (access, extraction, navigation, data cleaning, metadata, tools); and Needs. The codebook was then used for coding the interview data by three coders. A sample (22%) of all data was used to calculate intercoder reliability among

the three coders, yielding satisfying statistics, with Fleiss' Kappa = .963. A Kappa value between 0.81 and 1.00 would be considered almost perfect agreement (Fleiss, 1981).

3. Results

The following section presents the major themes that emerged from our data.

3.1. Workflow

To address RQ1, participants noted variability in their data search strategies, depending on the search objectives, data types, and cost. While each interviewee might implement a unique approach, their workflow followed similar trajectories, in which participants would connect a research question with data, attempt dataset searches, download and clean the datasets, evaluate the reliability of the data, and often repeat this process multiple times to find a usable dataset.

3.1.1. Connecting a Research Question with Data

Participants identified two key approaches to connecting research questions with data: question-driven (questions guide the data search) versus data-driven (data informs the questions). Data journalists and librarians tended to start with a question in mind. For them, a key challenge is creating the proper search query based on the question. Not knowing the proper terminology especially inhibits those working in interdisciplinary domains.

Engineers, data scientists, and researchers in natural sciences sometimes started with data first:

> Sometimes I'm just kind of brainstorming by looking around at the data, and investigating, maybe do some quick plots and see if I spot any interesting patterns or anything. And then sometimes I have an idea. (P6)

With such a data-driven approach, access to high-quality, cutting-edge datasets predetermined the success of their work.

3.1.2. Time-Consuming Process

Dataset search can be time-consuming and tedious, and can take a few hours to weeks or months, which became inhibitive to participants' work. Data journalists (e.g., P1 and P2) sometimes had to change their strategies to meet with publication deadlines if they could not locate or access the data they originally needed in time. They would usually pursue stories in which they knew 'the clearest path to victory already.'

The problem of search time is not limited to data journalists. P16 once spent four years searching for a specific dataset. P8 described dataset search as 'a whole research project on its own.' Sometimes, they had to reuse old datasets because they could not afford 'spending months curating a new dataset.'

3.1.3. Major Roadblocks

Another challenge that the participants experienced was to determine data usability. Half of the participants found it almost impossible to determine if a dataset was usable without first downloading it. Yet the download process can be complex. 19 of our 24 participants experienced difficulties in extracting data. These problems ranged from needing to take multiple steps to download, converting data out of PDF format, encountering paywalls or restricted data that requires screening, or having to download data piece by piece rather than all at once.

The next step, data cleaning, required various coding software and tools (e.g., Microsoft Excel, Python, SQL, Jupyter Notebooks, and machine learning more broadly), and thus posed challenges to the participants. Then, a variety of sense-making processes (e.g., data conversion, data categorization, data normalization) were adopted

to ensure that the data was indeed usable. If the dataset quality was unsatisfying, our participants had to start with a new search.

3.2. Source

Another theme emerged from the interview data is *source*. Finding a credible and reliable source is key for ensuring the effectiveness and efficiency of dataset search.

3.2.1. Governmental websites

Participants noted that they had 'go-to' sources. All but one of the interview participants mentioned the use of government agency websites for data acquisition. Data provided by government agencies tend to be domain-specific: an anthropologist (P4) used maps produced by Federal Emergency Management Agency, an astrophysicist (P6) was most familiar with a National Aeronautics and Space Administration database, and an electrical engineer (P14) used data published by the Department of Energy.

Though viewed as credible sources, governmental data sites were not without their flaws. A data journalist (P1) who usually gathered data from government agencies and departments noted:

> I go to the agency and might get data from the state but the data is city data, so you're saying, 'Is it accurate?' Other times, I've gotten data from open data sites and I've gone to the agency and said, 'We're writing about this, can you help me understand it?' and they've said, 'No, don't use that, it's not accurate.'

Participants reported high expectations of the governmental sites, as they considered government agencies as responsible for data openness and accuracy. Several of the participants suggested usability issues in these sites, as discussed in later sections. For these governmental sites, focusing only on experts or a selective group of users, rather than the general public, was described as a 'disservice' (P19).

3.2.2. Commercial websites

Google was another widely adopted source for data acquisition, especially among participants from fields such as computer science, data science, engineering, and physics. Google appeared as the most versatile commercial website, as participants used it to search for relevant literature (p10 & p20), image recognition and classification test sets (p8), and even mathematical formulas (p13).

3.2.3. Prior publications

Participants referred to prior publications as sources for existing or secondary data. Data journalists would strategically look for news articles that reported on the data of interest, which functioned as benchmarks for interpreting and presenting the data. Academic journals frequently served as data sources for researchers. Participants (P8, P10, P13) reviewed academic paper references and abstracts for 'properties of interest' (P13) and relevant information such as metadata or links to datasets (P14). P11 mentioned academic conference proceedings as the timeliest source for impactful work.

3.2.4. Data authors and experts

Participants relied heavily on direct communication to gain access to datasets as well as clarification and updates. Academic researchers would rely on collaborators and colleagues to obtain datasets or gain information about the datasets; all of the researchers cited collaboration and networking at conferences as an important source:

> A lot does happen in face-to-face conferences... Researchers will sometimes say: 'There's a link in the paper to our dataset,' or 'our data are available on Github (Author note: a Web-based hosting service).' (P3)

Data journalists would often contact the data authors or domain experts to obtain original datasets, inform the data authors about data use, or verify their interpretation of

the data. As P2 said: 'I will call up an academic who I know works in a certain area and say, "Do you know of anyone who studied this and may have some data on this?"' This is considered a more direct and reliable way of obtaining data than from a public data website.

Direct communication effectively facilitated access to data, but only occurred through existing social connections. Contacting unknown data authors was considered inappropriate, and most of the data sites provided little contact information and no alternative communication channels.

3.3. Technical Barriers

Ultimately, in response to RQs 2 and 3, dataset search remains a huge technical challenge to participants.

3.3.1. Searchability

Participants noted the lack of organization and clarity within databases and data websites, as the disorderly nature of the websites made it sometimes impossible to find relevant datasets. Participants then turned to the search box, but they (e.g., P10, P17, P21) had trouble generating an effective query, and many of the interface provided no assistance or query refinement. It was evident that there was a mismatch between users' thought processes for formulating queries and the queries that search engines required to produce desirable results.

Another usability problem was the lack of clarity within, and sometimes the absence of, dataset descriptions. Participants expected dataset descriptions to be thorough and correctly indexed by search tools. They also relied on the dataset descriptions and meta-data to gauge data relevancy. In reality, however, many websites offered very limited information: '[The] description of the dataset on the website wasn't

super transparent... a lot of times they didn't even describe what the columns were, what the fields were' (P3).

It was also difficult to 'refine results for individual research questions' (P4). Participants attempted to narrow their searches by specifying file formats and publication date, using filters, etc. However, these strategies might not be supported by the data sites, leaving users unable to specify search parameters and with search results 'comically random' (P1).

For researchers, the inability to refine searches made it impossible to identify data from 'cutting-edge research' (P4) or for niche areas, as search results normally are ranked by popularity. For data journalists, the lack of searchability meant that they had to abandon their 'aspirational stories' as a "cost-benefit trade-off—how important is this data to the story, and then how important is the story" (P1).

3.3.2. Interactivity

Interactivity affordances have been conceptualized as 1) user-interface interaction and 2) user-user interaction (e.g., McMillan & Hwang, 2002; Sundar et al., 2015); existing data sites lack both. As discussed in the previous section, participants could perform very limited actions on the data websites: Only two participants reported the ability to perform actions such as filtering or regrouping of data; only one participant (P6) noted the ability to visualize or compare datasets. Data sites were even more limited in their ability to foster interaction amongst users. Participants reported the inability to communicate with authors or other users via the sites. In this section, we will focus primarily on this CMC aspect of interactivity.

The majority of journalist participants noted their need to reach out to the data authors or domain experts who could help clarify the data in use, or other reporters and editors who might have worked with similar datasets; similarly, academic participants

would reach out to researchers or data managers with questions or to request raw data. The lack of user interaction affordances made such collaboration difficult or impossible. As a result, important questions were left unanswered and flawed datasets were left unchecked: P6 could not report an error to the data author, and P18 was unable to alert authors of duplicates in their datasets. They expressed their frustration over having 'no real venue' to establish a channel for communication if there was no existing social connection.

Several participants suggested adding a social component to the data sites which would connect users directly with the data authors, and more importantly, enable an open data community. The only site with a social feature that our participants (i.e., P1, P2, P23) suggested was data.world:

> data.world is open to everyone, makes it easy to upload your own data, is really simple to use such that the data producers can feel comfortable about uploading their information and having it shared. [Its social features] make uploading data and sharing it a joy, allow the data producers to see how their data is being used, and allow a kind of dialogue back and forth between data producers and data users. I think that both incentivizes data producers to want to be a part of the project, and then gives data users a place to meaningfully get feedback and advice. (P2)

Other participants suggested that interaction among data authors and data users could function as a way to *uphold data integrity*. On data.world, where users can 'upvote and discuss' a dataset, data authors are encouraged to enhance data quality (e.g., by updating and cleaning the data, adding annotations, and answering questions from other users). Such a mechanism can be adopted by data websites to ensure data quality:

> [Users] feel a lot more confident in a dataset if [authors] post their methodology when posting that dataset. I've uploaded only one dataset to data.world, and I exhaustedly tried to document how I arrived at that dataset and the other data sources I used. So, if other folks could apply a similar level of rigor, I think that

> would go toward somewhat alleviating the concern that anyone can post any data. And if there's a rating system [in which] people can flag a dataset that is poorly documented, I think that would help, too. (P23)

3.3.3. Usability

Usability refers to properties of a system that determine its ease of use, efficiency and effectiveness (Hearst, 2009; Shneiderman & Plaisant, 2004). Usability is dependent upon the dynamic interplay of the technological tool, its user, the task, and the environment (Shackel, 1991). In their evaluation, participants found few data websites to be 'usable,' or created with non-expert users in mind:

> [The data websites are] often not designed by people who are using the data. [The governmental agencies] have a mandate to put up data, [they'll] put data up instead of thinking, 'What do we think people need to be able to evaluate the government?...we should put them up in a way that's useful and understandable and... findable.' (P1)

A majority of the participants referred to Google advanced search as one of the usable tools. It provides various search parameters (e.g., exact keywords, file type, site or domain, region, etc.), which other sites do not offer. General search engines like Google, however, would not return datasets-only results, results specific enough to the topic, or allow users to parse pages adequately.

Half of our participants described instances in which they were unable to access data, due to paywalls (P5, P12), partial access to the dataset (P4, P19), the requirement of a user account affiliated with a university (P6), or other forms of governmental screening or restrictions (P1, P11, P14, P15).

The most reported usability problem was data extraction (i.e. dataset download and conversion). 19 participants reported difficulties in extracting data from a data website; they experienced multi-step download (P6, P21), dead links (P5, P12), and

software incompatibility (P8, P11). Dataset download and conversion may require programming skills, which most non-experts did not have. Alternatively, participants completed the procedure manually: P4 "physically looked up 700 [data points]", and P1 considered reformatting datasets as one of the most challenging aspects of their work. For data journalists in local newsrooms and researchers in underfunded institutions without adequate resources and training, this usability issue posed a great challenge. Participants attributed this inaccessibility to the lack of incentive or motivation, or incompetence, to make data files readily usable: 'Sometimes [the data managers] will have a table, print it, and they'll scan it, and then they just put that in a PDF, so that the whole thing is an image' (P24). Participants observed this across various governmental agencies, institutions, and in their international work experiences.

Participants also found the data quality to be undesirable. They had to spend a significant amount of time cleaning the data, e.g., correcting column titles, searching for missing data, or checking number format. Online data were often out of date. For instance, government data repositories took years to publish federal data; institutional data centers and major search engines such as Google did not reflect 'the most up-to-date work being done' (P4), either. To enlist the most up-to-date data and encourage collaboration, P4 attempted to create a crowdsourcing site that would be an interactive repository for researchers to share their work. However, without a culture or a community supportive of data sharing, few people used the site:

> How is it that everybody is so worried about where other people are doing their research and now we build a tool for them to be able to do that and they can't even be bothered to enter their information.

Emerging data websites like this one dealt with problems such as lack of traffic, technical support, and attention paid to the growing need among non-expert users for

cross-discipline datasets.

3.4. Summary of Findings

The interviews revealed three primary problems that non-expert users encounter in data searches: *searchability*, *interactivity*, and *usability*.

Searchability. A foundational issue for non-experts was the lack of organization and searchability within databases. Considering that 'Web search engines are designed for documents, not data,' and that 'contextual or personalized results are practically non-existent for data search' (Koesten, Kacprzak, Tennison, & Simperl, 2017), users often had a difficult time finding relevant datasets. Ineffective organization and inadequate descriptions slowed and disrupted the whole workflow.

Interactivity. Most data sites were structurally inventories of datasets and offered few interactive features, if at all, for users to perform actions on an interface or communicate with a data author or another user. Social and communication functions were in particular limited. Likening a dataset to a source, participants demanded channels of communication to ensure accuracy and determine validity.

Usability. Existing data websites were not designed with non-experts in mind. Usability and user-friendliness were lacking across different data search tools and domains, making search processes overly taxing. Government agencies and other entities were mandated to publish data but without standards or quality control, which limited data use and dissemination.

Participants suggested functionalities and features to enhance searchability, interactivity, and usability: e.g., keyword search (71% of all participants); data description coherence and standardization (63%); easy, direct download (58%); transparency (58%); improved meta-data (50%); user-generated tagging (46%); direct

communication among data stakeholders (46%); and visibility of recent or specialized data (46%).

4. Discussion

Publicly available data has the immense potential to contribute to interdisciplinary scientific discovery (Jeppesen, Ebeid, Jacobsen, & Toftegaard, 2018), journalistic innovation (Lesage & Hackett, 2014), effective policymaking (Napoli & Karaganis, 2010), and better citizenship (Carmi, Yates, Lockley, & Pawluczuk, 2020). At the same time, studies (e.g., Bertot, Jaeger, & Grimes, 2010; Janssen, Charalabidis, & Zuiderwijk, 2012; Kitchin, 2014; Poel, Meyer, & Schroeder, 2018) have indicated, as observed in our study, that many issues (e.g., the lack of accessibility and openness, the problems with literacy and usability, the barriers for adoption, participation and collaboration) have created a gap between the promised benefits and the present challenges. Our research contributes to the existing literature by providing empirical evidence of the issues in non-expert data searches and offers insights into ways to narrow the gap between expectations and reality. Our findings show that existing data repositories and websites are usually siloed by discipline or domain, more fundamentally driven by funding and proprietary or institutional structures. Non-experts frequently experience terminological, technical, financial and authorization difficulties for cross-domain data access and use.

4.1. Understanding non-expert users

One of the fundamental problems is that few data websites or search tools are designed for non-experts. Data sources (such as government agencies and companies) may deprioritize, or even intentionally resist, data access and usability for the general public. Little research on dataset search has examined the search patterns and preferences of

non-expert users, nor has ‘non-expert’ been clearly defined or conceptualized in the context of dataset search. As shown in our study, data systems and interfaces developed for experts and experienced users may not be readily usable for non-experts, due to their distinctive levels of experience and knowledge. Non-experts find existing tools too complicated and become prohibited from utilizing them, hence creating a ‘Big Data Divide’ (Andrejevic, 2014). To enhance the usability of dataset search tools, data providers and interface designers need to better understand the capabilities, processes and work environments of the non-expert users.

4.2. The complex problem of searchability

Critical for the effectiveness of a dataset search tool targeting non-experts is its searchability. Non-experts rely heavily on a clear organization of the data content with standardized, concise and transparent data descriptions or metadata. They expect the interface to be highly searchable, by providing logical indices, or highly functional search engines with search parameters to narrow down search results. These features facilitate users to locate datasets that are *available*, but not necessarily *usable*. Participants described the tedious process of downloading and cleaning the dataset and then evaluating its usability. This indicates that users can benefit from a direct preview of the dataset content or searching dataset at a cell-level. To assist quality assessment, the interface could implement metrics indicating the popularity or impact of a dataset (e.g., recently downloaded, most downloaded, most searched for, etc.).

4.3. Need for communication

This study also adds to the existing literature by revealing the communication needs of non-expert users. Data journalists and researchers constantly seek and benefit from direct contact with data authors and experts. When communication channels are

unavailable on data sites, users are forced to depend on their personal connections. Laypersons without such connections or networks have no means to seek help and may rely on unreliable sources. Social components (e.g., email, discussion board, subscription to data updates) built into these interfaces can decrease inhibition about contacting data authors and minimize the cost of establishing connections. While there are concerns over user-generated datasets (e.g., data quality, ownership, motivation, methodology), social mechanisms can help verify and vet the datasets, promote data reuse and collaboration, and encourage a deep, comprehensive understanding of data.

4.4. Constraints or liberation?

Digital technology has a significant influence on an individual's ways of knowing and doing, as well as an organization's norms and values (Lewis & Westlund, 2015). In emerging fields such as data journalism and interdisciplinary research where technological advancements drives and defines its development, dataset search tools both enable and constrain innovation. The question researchers and journalists ask themselves will become one of ethics and principles: if they can only obtain or utilize summary, second-handed, or outdated data, how can their work be of quality, significance, and truth? How can scholars generate 'transdisciplinary' research (Aboelela et al., 2007), and how can journalists serve the news media's 'fourth estate' function (Felle, 2016)? How will they challenge the boundary of their communities, contribute their utmost effort to the public good, and truly fulfil their purpose in society?

Technology tools significantly influence non-expert data use and shape their decision-making and work quality. To facilitate non-experts' access and adoption of data, data websites and search tools should implement affordances crucial for liberating non-expert users from the barriers for 'discovering, annotating, comparing, referring,

sampling, illustrating and representing (Unsworth, 2000)' data in their work. Appendix B summarizes key factors and design guidelines for supporting non-expert data search.

4.5. Broader implications

This study has two broader implications. First, public data cannot contribute to a better-informed society without effective dataset search tools. Second, barriers for non-expert data use will likely lead to misinformation and inequalities. Both relate to the fact that, primarily focusing on expert users and specialized domains, prior research and interface design often overlook the data use of a wider range of user population, with a vast range of needs, purposes, skills and resources, and the implications of the existing challenges and barriers. Our study defines and investigates non-expert dataset search, and motivates future research to rethink about users of public datasets.

The 'Big Data promise' states that the more data you have, the more effective policymaking becomes, and the more likely a substantive social change can be created. However, this could only be true when citizens are empowered to participate in data practices and when barriers such as literacy, usability, accessibility and functionality are brought down (Bertot et al., 2010). While data literacy is a long-term goal, improving usability and accessibility of data services should an immediate goal. Being non-expert users' 'go-to' data source, government agencies, in particular, have the normative responsibility to make data not only available but also usable to the general public.

Allowing non-experts to access and utilize reliable data can also help combat misinformation and disinformation. Search engines are now important information sources and define the nature and diversity of the content received by the users (Steiner, Magin, Stark, & Geiß, 2020). Expertise or technical barriers could change the informational landscape between expert and non-expert users. Exposing individuals to relevant datasets, enabling fact-checking, and connecting them to credible sources will

serve as mechanisms to escape an echo chamber and to debunk misinformation. To sustain a democratic society, 'public authorities need to make sure that all citizens have equal access and can easily use [open government data]' (Wirtz, Weyerer, & Rösch, 2019) by designing their platforms to meet the needs of all sociodemographic populations and communities.

4.6. Limitations

Our study has several limitations. First, it involved a relatively small number of participants. It is particularly challenging to find data journalists who are available to participate in our study. However, data journalists and researchers provided similar yet unique perspectives, as the two groups operate with different time frames, different ranges of topics, and different skillsets and social connections. Future research can collect data from a larger sample with a wider range of use cases and scenarios, which would allow a more in-depth analysis of different characteristics, such as users' skills, training, experience, access to tools and resources, as well as demographic and socio-economic factors, in relation to their use, preferences and expectations of dataset search tools. Input from fields and domains beyond academia and newsrooms will contribute to a better understanding of the data needs and purposes of non-expert users, and how dataset search tools can facilitate the various tasks, practices and workflows.

5. Conclusion

This study explores the ways through which researchers and journalists approach dataset search problems, how they apply current technology to solve these problems, and the challenges and barriers that they encounter in their work with data. These non-expert users represent a rapidly growing population who are starting to adopt and use data search tools. However, with technological limitations of data search tools in the

aspects of searchability, interactivity and usability, modern data-driven production for non-experts remains a challenge. We have found that little attention has been paid to non-experts' emerging data needs, leading to significant constraints to the design and development of technological tools for supporting non-expert users. Our findings underline the critical impact of the design, development and deployment of technological tools to enable the meaningful use of today's increasingly available data toward a civil society.

6. Acknowledgement

This work was supported by the National Science Foundation under Grant No. III-1816325.

References

Aboelela, S. W., Larson, E., Bakken, S., Carrasquillo, O., Formicola, A., Glied, S. A., . . . Gebbie, K. M. (2007). Defining interdisciplinary research: Conclusions from a critical review of the literature. *Health Services Research*, *42*(1), 329–346.

Al-Maskari, A., & Sanderson, M. (2011). The effect of user characteristics on search effective ness in information retrieval. *Information Processing & Management*, *47* (5), 719–729. Andrejevic, M. (2014). Big data, big questions—the big data divide. *International Journal of Communication*, *8*, 17.

Baack, S. (2018). Practically engaged: The entanglements between data journalism and civic tech. *Digital Journalism*, *6*(6), 673–692.

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government information quarterly*, *27*(3), 264–271.

Bonaque, R., Cao, T. D., Cautis, B., Goasdou´e, F., Letelier, J., Manolescu, I., . . . Thomazo, M. (2016). Mixed-instance querying: a lightweight integration architecture for data journalism. In *Vldb.*

Borges-Rey, E. (2016). Unravelling data journalism: A study of data journalism practice in British newsrooms. *Journalism Practice*, *10*(7), 833–843.

Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). OPEN—Enabling non-expert users to extract, integrate, and analyze open data. *Datenbank-Spektrum*, *12*(2), 121–130.

Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, *9*(2), 1–22.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Coddington, M. (2015). Clarifying journalism's quantitative turn. *Digital Journalism*, *3*(3), 331-348.

Cushion, S., Lewis, J., & Callaghan, R. (2017). Data journalism, impartiality and statistical claims: Towards more independent scrutiny in news reporting. *Journalism practice*, *11*(10), 1198–1215.

Duggan, G. B., & Payne, S. J. (2008, April). Knowledge in the head and on the web: Using topic expertise to aid search. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 39-48).

Espinosa, R., García-Saiz, D., Zorrilla, M., Zubcoff, J. J., & Mazón, J. N. (2013, August). Enabling non-expert users to apply data mining for bridging the big data divide. In *International Symposium on Data-Driven Process Discovery and Analysis* (pp. 65-86). Springer, Berlin, Heidelberg.

Felle, T. (2016). Digital watchdogs? Data reporting and the news media's traditional 'fourth estate' function. *Journalism*, *17*(1), 85–96.

Fink, K., & Anderson, C. W. (2015). Data Journalism in the United States: Beyond the 'usual suspects'. *Journalism studies*, *16*(4), 467–481.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. 2nd edn. New York, NY: John Wiley & Sons.

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory: Strategies for qualitative research. *Sociology Press*.

Gray, J., Chambers, L., & Bounegru, L. (2012). *The data journalism handbook: How journalists can use data to improve the news*. O'Reilly Media, Inc.

Hearst, M. (2009). *Search user interfaces*. Cambridge University Press. Holman, L. (2011). Millennial students' mental models of search: Implications for academic librarians and database developers. *The journal of academic librarianship*, *37*(1), 19–27.

Heflin, J., Davison, B. D., & Jia, H. (2021). Exploring Datasets via Cell-Centric Indexing. In *DESIRES* (pp. 53-60).

Hölscher, C., & Strube, G. (2000). Web search behavior of internet experts and newbies. *Computer Networks*, *33*(1-6), 337–346.

Jahng, M. R., & Littau, J. (2016). Interacting Is Believing: Interactivity, Social Cue, and Perceptions of Journalistic Credibility on Twitter. *Journalism & Mass Communication Quarterly*, *93*(1), 38–58.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258–268.

Jenkins, C., Corritore, C. L., & Wiedenbeck, S. (2003). Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. *IT & Society*, *1*(3), 64-89.

Jeppesen, J. H., Ebeid, E., Jacobsen, R. H., & Toftegaard, T. S. (2018). Open geospatial infrastructure for data management and analytics in interdisciplinary research. *Computers and Electronics in Agriculture*, *145*, 130–141.

Kelly, D., & Cool, C. (2002, July). The effects of topic familiarity on information search behavior. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 74-75).

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage.

Koesten, L. M., Kacprzak, E., Tennison, J. F. A., & Simperl, E. (2017). The trials and tribulations of working with structured data: A study on information seeking behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1277–1289).

Lesage, F., & Hackett, R. A. (2014). Between objectivity and openness—the mediality of data for journalism. *Media and Communication*, *2*(2), 42–54.

Lewis, S. C., & Westlund, O. (2015). Big data and journalism: Epistemology, expertise, economics, and ethics. *Digital journalism*, *3*(3), 447–466.

Lofland, J., Snow, D., Anderson, L., & Lofland, L. (1971). Analyzing social situations: A guide to qualitative observation and analysis. *Belmont, CA.: Wadsworth*.

Loosen, W., Reimer, J., & De Silva-Schmidt, F. (2020). Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016. *Journalism*, *21*(9), 1246–1263.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33.

Mao, J., Liu, Y., Kando, N., Zhang, M., & Ma, S. (2018). How Does Domain Expertise Affect Users' Search Interaction and Outcome in Exploratory Search?. *ACM Transactions on Information Systems (TOIS)*, *36*(4), 1-30.

McMillan, S. J., & Hwang, J.-S. (2002). Measures of perceived interactivity: An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *Journal of Advertising*, *31*(3), 29–42.

Napoli, P. M., & Karaganis, J. (2010). On making public policy with publicly available data: The case of US communications policymaking. *Government Information Quarterly*, *27*(4), 384–391.

Nguyen, V., Rybinski, M., Karimi, S., & Xing, Z. (2022). Search like an expert: Reducing expertise disparity using a hybrid neural index for COVID-19 queries. *Journal of Biomedical Informatics*, *127*, 104005.

Noy, N., & Brickley, D. (2017). Facilitating the discovery of public datasets. *Google Res. Blog post. Available online at: https://research.googleblog.com/2017/01/facilitating-discovery-of public.html*.

Poel, M., Meyer, E. T., & Schroeder, R. (2018). Big data for policymaking: Great expectations, but with limited progress? *Policy & Internet*, *10*(3), 347–367.

Schroeder, R. (2020). Big data and cumulation in the social sciences. *Information, Commu nication & Society*, *23*(11), 1593–1607.

Shackel, B. (1991). Usability–context, framework, definition, design and evaluation. In B. Shackel, & S. J. Richardson (Eds.). *Human Factors for Informatics Usability*, 21–37.

Shneiderman, B., & Plaisant, C. (2004). *Designing the user interface: Strategies for effective human-computer interaction*. 4th edn. Boston: Pearson/Addison-Wesley.

Slone, D. J. (2002). The influence of mental models and goals on search patterns during web interaction. *Journal of the American society for information science and technology*, *53*(13), 1152–1169.

Steiner, M., Magin, M., Stark, B., & Geiß, S. (2020). Seek and you shall find? a content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society*, 1–25.

Stoneman, J. (2015). Does open data need journalism? (working paper). *Reuters Institute for the Study of Journalism, Oxford University*.

Suhr, B., Dungl, J., & Stocker, A. (2020). Search, reuse and sharing of research data in materials science and engineering—A qualitative interview study. *PloS one*, *15* (9), e0239216.

Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME) four models for explaining how interface features affect user psychology. *The handbook of the psychology of communication technology*, 47–86.

Tremayne, M., & Dunwoody, S. (2001). Interactivity, information processing, and learning on the World Wide Web. *Science Communication*, *23*(2), 111–134.

Umemoto, K., Yamamoto, T., & Tanaka, K. (2016, April). How do users handle inconsistent information? the effect of search expertise. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (pp. 1066-1071).

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. In *Symposium on Humanities Computing: Formal Methods, Experimental Practice. King's College, London* (Vol. 13, pp. 5–00).

White, R. W., Dumais, S. T., & Teevan, J. (2009, February). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 132-141).

Wirtz, B. W., Weyerer, J. C., & Rösch, M. (2019). Open government and citizen participation: an empirical analysis of citizen expectancy towards open government data. *International Review of Administrative Sciences*, *85*(3), 566-586.

Wongsuphasawat, K., Liu, Y., & Heer, J. (2019). Goals, process, and challenges of exploratory data analysis: an interview study. *arXiv preprint arXiv:1911.00568*

Young, M. L., Hermida, A., & Fulda, J. (2018). What makes for great data journalism? A content analysis of data journalism awards finalists 2012–2015. *Journalism Practice*, *12*(1), 115–135.

Zahidi, Z., Lim, Y. P., & Woods, P. C. (2014, August). Understanding the user experience (UX) factors that influence user satisfaction in digital culture heritage online collections for non-expert users. In *2014 Science and Information Conference* (pp. 57-63). IEEE.

Zelenkauskaite, A., & Sim˜oes, B. (2014). Big data through cross-platform interest-based inter activity. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)* (pp. 191–196).