Bitacora: A Toolkit for Supporting Non-Profits to Critically Reflect on Social Media Data Use

Adriana Alvarado Garcia
adriana.ag@ibm.com
IBM Thomas J. Watson Research Center
Yorktown Heights, NY, United States

Marisol Wong-Villacres
lvillacr@espol.edu.ec
Escuela Superior Politécnica del Litoral
Guayaquil, Ecuador

Benjamín Hernández
NVIDIA
Santa Clara, CA, United States

Christopher A. Le Dantec
c.ledantec@northeastern.edu
Khoury College of Computer Sciences, Northeastern
University
Boston, United States

Abstract

In this paper, we describe the design and evaluation of the toolkit Bitacora, addressed to practitioners working in non-profit organizations interested in integrating Twitter data into their work. The toolkit responds to the call to maintain the locality of data by promoting a qualitative and contextualized approach to analyzing Twitter data. We assessed the toolkit's effectiveness in guiding practitioners to search, collect, and be critical when analyzing data from Twitter. We evaluated the toolkit with ten practitioners from three non-profit organizations of different aims and sizes in Mexico. The assessment surfaced tensions between the assumptions embedded in the toolkit's design and practitioners' expectations, needs, and backgrounds. We show that practitioners navigated these tensions in some cases by developing strategies and, in others, questioning the appropriateness of using Twitter data to inform their work. We conclude with recommendations for researchers who developed tools for non-profit organizations to inform humanitarian action.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Field studies**; *Interaction design process and methods.*

Keywords

social media data, user-generated content, NGO, non-profit, humanitarian context, data work, data experts, data annotation, ground truth, toolkits, design

1 Introduction

The push for technological innovation in the non-profit, humanitarian sector is increasingly championing new ways of harnessing social media that go beyond the traditional practices of spreading information about non-profit work and needs [102] or building connections with volunteers, community, and media [19]. Specifically, motivated by existing research depicting social media data as a promising pathway for efficiently addressing on-the-ground crisis [3, 10, 21, 52, 56, 74, 82, 114, 120], funders and other organizations promote social media data-centered interventions as a desirable mechanism to ensure non-profits invest resources optimally [11, 21, 25, 74]. However, recent work on Critical Data studies and Human-Computer Interaction (HCI) has stressed the complexities entailed in mobilizing social media data from its site of production to non-profits' operations and goals [2, 4, 5, 21, 24, 30, 93, 112]. Such a process demands non-profits to have staff knowledgeable in data analysis, technological resources, and time, all of which are hard to secure for these organizations [7, 19, 67, 102]. Further, mobilizing qualitative data from its site and time of production to a completely different context, strips it out of meaning [2, 4, 21, 76], making the derived lessons dangerous to apply in on-the-ground humanitarian problems. Technological innovation in the form of social media data applications, thus, poses a critical challenge for humanitarian and non-profits; these organizations must learn to evaluate when and how it is appropriate and productive to use social media data without falling into techno-centric innovation discourses about these data's actual capabilities for work on the ground [2, 5, 11, 16, 20, 37, 60, 65, 68].

The design and use of toolkits that could support non-profits as they critically evaluate their social media data explorations emerge from social innovation and HCI literature as a promising alternative to this problem [18, 38, 45, 54, 62, 81, 96, 103, 106, 107]. In the context of community and social innovation, researchers and practitioners have proposed toolkits to guide community work as well as social impact initiatives, facilitate the use of design thinking methodologies, and protect human rights [38, 54, 62, 96, 103]. In regards to the field of HCI, for decades the development of toolkits has played an important role to produce *"generative platforms"* [46] that can "inform people about technologies" [103] and "best practices" [81] for a specialized situation in a "streamlined, scripted fashion" [95]. As "curated collections of tools and materials" [119],

toolkits can scaffold deliberation processes by translating complex issues into concrete and accessible technologies development and use practices [81, 95].

Existing studies on the design of toolkits for guiding technology-savvy users as they apply novel technologies such as Machine learning (ML), Artificial Intelligence (AI), and cybersecurity systems have stressed, however, that using toolkits to support responsible technology adoption is not free of problems [32, 97, 119]. Specifically, toolkits' goal of being a guide for best practices and representing a form of "professional vision" tends to provide decontextualized guidance [119], flatten nuances, limit perspectives outside of the tools within the toolkit[63, 81], and promote a solutionist perspective of social problems [119]. Further, toolkits' attempts to promote critical views amongst industry actors have not been proven effective yet either: due to time constraints, lack of organizational support, and poor alignment with working pipelines, industry actors do not find them usable or useful in practice [32]. Against this backdrop, it remains critical to explore the role of toolkits—if any—in slowing down technological innovation for non-technology-related organizations, such as humanitarian non-profits, and prompting them to critically decide, how, and if to take a technological resource out of its context for impacting communities on the ground.

In this paper, we address this critical need by describing the first author's situated experience designing and testing the *Bitácora* toolkit[1] and its proposed methodology. Bitacora's main goal was to prompt humanitarian and non-profit organizations that are considering computational uses of social media data to critically reflect on the role that Twitter data can have for supporting their operations. Throughout a fellowship and internship that together lasted 17 months, the first author worked with the Accelerator Lab Mexico, which belongs to the Accelerator Lab Network of the United Nations Development Programme (UNDP), the United Nations' lead agency on international development [1]. The first author worked with the Accelerator Lab Mexico in developing a critical and situated approach to use social media data as a source for actionable evidence informing crisis response. Following [103], the toolkit and its evaluation process with staff from the Accelerator Lab Mexico and two Mexican non-profits, one of them with no experience with social media computational analysis, emerged as an attempt to support non-profits' staff, which often holds different degrees of technological savviness, to autonomously engage in critical and responsible social media data explorations. This work describes the design intentions and decisions behind the Bitácora toolkit as an artifact that prioritizes a situated data analysis for slowing down innovation in the context of humanitarian and non-profit organizations. Further, it reports non-profit practitioners' reactions to the toolkit and the different ways in which these interactions shaped practitioners' perspectives on social media data as a technological innovation.

We found that the toolkit's design aspects across its proposed methodology generated three different reactions from practitioners: (1) at early stages of the methodology, the toolkit's request to use constraints such as keywords and time frames motivated practitioners to engage in deeper explorations of the problem; (2) as practitioners progressed in the methodology, they tended to adjust

their goals and notions of validity to Twitter's limitations rather than reflecting on the implications of their decisions; and (3) the request for a situated, qualitative analysis at the last stages of the process did momentarily slow down practitioners' pragmatic expectations on social media data but, in line with [32], it also clarified the difficulties of adopting this artifact in the long run given the amount of time, labor, and care it demands out of practitioners and non-profits.

As such, this work offers three contributions to researchers working to support non-profit organizations as they grapple with the push for technological innovations in the humanitarian sector. *First*, the toolkit's design, which emphasizes situatedness and seeks to support practitioners in challenging a reductionist rhetoric behind social media data adoption. *Second*, an analysis of how non-profits' staff members reacted to the toolkit-based intervention and reflected on their perspectives on Twitter data as an appropriate source of information. *Lastly,* a discussion on the possibility for a toolkit such as Bitácora to slow down social media data innovation, including (1) the design aspects that were successful in driving an in-depth reflection about social media data use; (2) feasible design pathways for ensuring that the toolkit can have a long-term impact in non-profits' use of social media data.

2 Related Work

2.1 Non-Profits and Technological Innovation

The discourse of technological innovation is often understood as the introduction of modern technologies that can disrupt, transform and overall, revolutionize how an organization operates [16, 31, 64, 66]. As such, it offers non-profit organizations numerous potential benefits directly connected to the generation of value, economic growth [58, 66], efficiency, empowerment, and overall, survival [16, 49, 57]. In particular, non-profits and their stakeholders envision technological innovation as critical for optimizing their administrative, service and marketing areas, communicating their services and mission to communities and funders, staying in touch with constituents, recruiting volunteers, raising funds, and assessing impact [16, 49, 57, 84]. However, as [16, 49, 58, 115] explain, the promises behind new technologies also feed the illusion that technological interventions are an easy-to-pursue imperative that "just needs" easy fixes to take place (e.g., an Internet connection, a database software, personal devices, training, or an IT consultant). These authors explain how the reality is quite different; pursuing technological innovation is a considerable undertaking that heightens the scrutiny that non-profits are under, entails greater demands for them—often with the same or fewer resources, and encourages a culture of competition, all of which is often irreconcilable with their reality and goals.

Non-profits operate within—and depend on—the power dynamics of a complex network of government systems and public policy fields [17, 70, 101, 115]. The transformational narrative behind technological innovation feeds into these dynamics, creating various axes of demands that are hard to satisfy and often in tension with each other. Non-profits are expected to "keep up" with the private sector progress [49], follow public sector and funding agencies' digital use recommendations (e.g., keep electronic records, set up online application processes, and provide impact and performance

---

[1]The word *Bitácora* means log book in Spanish.

data) [16, 51, 115] while also adopting technologies that align with the technology practices of the communities they serve [17, 84, 115]. Implementing technological change is also quite difficult due to non-profits' extraordinary constraints for securing and mobilizing financial resources, time, expertise, and clarity about the role of technology in their operations [16, 25, 37, 70, 80, 84, 115]. New technology, then, often enters this space in a improvised fashion (e.g., as part of a grant or "created on the fly"), with little understanding from the organization staff of how it connects with their mission [84, 85] and no clear direction on how to connect the innovation to the organizations' existing patchworks of information systems and equipment [16, 115]. As a result, non-profits often end up treating technological innovations as a commodity rather than an innovation enabler (e.g., using email as a communication tool rather than as a strategic tool for collecting email addresses or distributing electronic newsletters) [22, 26, 44] and struggling to maintain technological additions (e.g., facing versioning issues, data redundancy, and data inaccessibility [70, 84, 115]).

Against this backdrop, an important body of work has explored technology (re) design as a desirable pathway [37, 109, 110, 116, 117]. Some of the approaches suggested are to identify and leverage non-profits' resources (e.g., volunteers, small grants, community-guided initiatives) [37, 117], intervene in non-profits' technology adoption process [16, 47], and provide methodological techniques for long-term technology planning (e.g., technology inventory assessment) [84]. However, increasingly, research suggests that the path forward lies in a deeper reflection about the underlying logics and assumptions behind technological innovations for non-profits. In particular, this body of work argues that in prioritizing economic growth, the discourse of technological innovation promotes tools and systems that only care about turning non-profits in more efficient institutions, disregarding these organizations' goals, values, and expertise along the way [34, 49, 50, 64]. For example, [16] found that, datafication initiatives promote donors' need to scrutinize non-profits and erode non-profits' autonomy to design their data strategy and make meaning of the analysis. Likewise, in working to maximize the number of donors, charity-based technologies dismiss the care work needed for donors to develop a longer-term, trusting relation with non-profits and their missions [50]. As [49] explains, the rhetoric of innovation creates technologies such as automated birthday notes, digitized volunteering work, and customizable reports, that demand new practices and, as a byproduct, generate more overhead for non-profits. The technological demand and offer cycle, [49, 58, 115] warn, only replaces old remedies with newer, more efficient versions, but does not really offer a transformative approach to non-profits' systemic problems. Together with [9, 16, 72], these scholars encourage researchers to resist the technological innovation narrative that pushes for impossible futures and work with non-profits to offer them the support they need for rather disrupting innovation. This can include offering political support for non-use, intervening in the crafting of technological narratives, and working with non-profit actors to analyze technological interventions' overpromises. *Our research seeks to explore how to design such supporting mechanisms for international and humanitarian non-profits facing the pressure for engaging in social media data analysis as a desirable innovation.*

2.2 The Humanitarian Potential and Limitations of Social Media Data

Increasingly, important organizations and actors in the humanitarian field—including the World Humanitarian Summit [11, 74], external funders [25], and digital humanitarians [21] promote the datafication of the humanitarian non-profit sector. Datafication is indeed promising for helping non-profits to optimize their operations [25, 37, 60, 94] and make "more productive and empowered decisions" [15]. However, technological barriers [37, 47, 117], an incipient culture around strategic data use [70, 80, 84], and the pressure of funding agencies to define the structure of the data that needs to be collected [16, 27] often hinder organizations' ability to pursue this type of innovation. As [16, 111] explain, these factors often create a cycle of disempowerment that ends up flooding organizations with data that they cannot effectively use.

Research in Crisis Informatics and the Humanitarian sector has shed light on the potential for user-generated data—such as social media posts and comments—as a valid alternative: social media platforms are a low-cost, quick source of large amounts of data that grow by the second and are rich in expressed needs and opinions about conditions on the ground [3, 10, 21, 52, 56, 74, 82, 114, 120]. Further, due to its heterogeneity, social media data holds the promise of giving non-profits the freedom to explore the problems they need, without depending on stakeholders' demands [71]. The analysis of these data can, thus, help humanitarian non-profits to prepare for and react to crises by providing insights on disease outbreaks, potential risks, communities' resources, casualties and damages, and donation efforts [10, 40, 55, 59, 83, 121]). Social media data can also inform non-profits about the audience consuming their content, and generate evidence for funders about their impact and community connections [7]. However, the use of social media data is not free of challenges. One of the most evident barrier is that of access; as with any technological innovation, non-profits struggle to secure the financial, technological, and human resources needed to make use of social media analysis tools [7, 19, 67, 102]. [7] found that out of 20 non-profit actors, only 37% had used these tools to make sense of social media data, 30% had not knowledge that these tools existed and 50% did not have clarity on how to use these platforms to add value to their work. As such, most non-profits in the humanitarian sector only use social media for information dissemination [7, 71].

Another challenge is social media data's incompleteness and, thus, inability to provide complete and objective inferences about conditions on the ground [2, 4, 5, 21]. As Science and Technology Studies (STS)' scholars have increasingly argued, data analysis' dependence on human interpretation and socio-cultural values invariable produces incomplete lessons [11, 68, 105]. For example, techniques for classifying tweets based on keywords (e.g., lexicons related to an event) or characteristics (e.g., only those referring to situational updates), depend on a previous definition of the keywords or characteristics that are valid, which in turn, can fail to capture a considerable fraction of relevant portions of data [41, 53, 75, 98]. Machine learning techniques also face this issue: the "dependent variable" needed to train and validate ML models called ground truth data also demands human decisions on the right representation of data [8, 91]. Data classification processes, thus, invariably

produce an over-synthesized version of reality that "becomes naturalized as 'the way things are'"[33, 35] while dismissing complex experiences behind (e.g., groups that are not present in the collected data such as older citizens, citizens without cellphones living in rural communities, citizens who prefer a different social media platforms, etc.) [7, 86, 87, 90, 123]. In addition to generating incomplete inferences, social media data analysis runs the risk of producing highly misguided insights [2, 93, 112]. When non-profits use these data, for example, they do so without considering the context where data was originally produced [24, 30]. As a result, it becomes hard for these organizations to determine what part of these data is false, non-factual, outdated or just irrelevant [40]. Given the high volume and velocity of online information, manual processing is not an option either [40]. As [20] and [76] found in their studies of crisis relief efforts, losing such valuable contextual richness can lead to insights about communities that do not align with how communities perceive themselves and end up "silencing local experts."

For these reasons, it is essential for non-profit organizations to understand these data limitations before they look for value in it; as [68] explains, "instrumentation alone is sufficient to deliver value". For example, organizations looking for "compelling examples reflecting the alleviation of suffering through [their] assistance programs" would benefit from understanding that those might not be easy to find or filter out [76]. Likewise, organizations trying to challenge critical issues such as institutionalized oppression need to know that social media data might not provide the verifiable evidence they need [2, 5]. Following the reflections of [11, 20], we argue that it is critical to support non-profit organizations in their process of defining what is valuable about social media data. This can entail facilitating reflections on why they collect one set of data vs others, how they consider—or not—data's production context, how they negotiate data, and how they make connections from data to action [11, 16, 20, 60, 65]. Our research is an initial attempt to advance such support via a toolkit that elicits critical reflections on social media data for non-profit practitioners.

2.3 Toolkits and Critical Perspectives on Technology

In its most general definition, the term toolkit describes a "set of tools that are used to making or repairing something", and "skills and knowledge that are useful for a particular purpose or activity" [95]. Across time, however, different disciplines have repurposed the term, using it to define a set of physical tools (e.g., the first-aid kit) as well as tools for thinking and decision-making (e.g., social justice kits) [95]. As [81] argues, in their essence, toolkits refer to scripted "best practices" for a particular situation, "a modular grouping of different perspectives, methods, and interrogatories" that formalize ways to approach the world [36]. In the field of HCI, the term has mostly referred to software and hardware components that make problems in the field (e.g., prototyping, designing, and developing interactive computing systems) easier to approach[78] As the ethical implications of new technologies have become more visible and debatable, the field has increasingly shifted to a definition of toolkits more similar to that coined by [36, 81]: as epistemological frameworks that provide technology adopters with perspectives, methods, and interrogatories for approaching the complex issues

created by technological innovations [18, 95, 100, 103, 104]. In the cybersecurity space, for example, toolkits such as the Access Now's Digital Security Helpline, and the SecureDrop's anonymous transfer service, leverage human support to ease non-technological users' engagement with security protection practices [97]. Other toolkits rather take the shape of tutorials or step-by-step guides for instructing the users how to improve security protection practices. In the context of technology design, researchers and practitioners have proposed toolkits for engaging diverse—sometimes marginalized—groups in the practice of design [18, 95]. The Cambridge Inclusive Design toolkit, for example, supports the creation of technologies that are "are accessible to, and usable by, as many people as reasonably possible"[45]. Toolkits such as the Equity Centred Community Design and the Community Led Co-Design Toolkit explicitly center on supporting communities as they engage in design work [95], and some, like the Building Utopia toolkit applies non-Western tenets—Afrofuturism—to elicit a different way of understanding and envisioning the future [18].

Another domain where toolkits-as-epistemological-frameworks are quite prevalent is the ethical adoption of AI solutions [32, 106, 107, 119]. While these toolkits all seek to promote ethics in AI design and development, they greatly vary in the type of approach they use. Microsoft's Fairlearn toolkit [14] and IBM's AI Fairness 360 toolkit [12], for example, provide timelines for industrial practitioners to "understand, assess and mitigate machine biases" [107]. To do this, these toolkits offer practitioners ready-to-use fairness metrics and bias mitigation algorithms that aim at facilitating practitioners' decision-making [124]. Other toolkits, such as the Model Card Authoring Toolkit encompasses reflection-oriented artifacts to rather prioritize practitioners engaging in debate, discussion, and negotiation before choosing "the right model" for the task at hand [107]. Google products seek to take critical reflection one step further by offering toolkits such as "What-if" and the ML-fairness-gym which helps practitioners to visualize the impact of fairness metrics [118], and even simulate the potential long-term social impact of machine learning-based decision [29].

Despite such emphasis on eliciting reflection, recent in-depth analysis of the role that toolkits play in guiding users as they engage with technological innovations suggest that, instead of eliciting critical reflection and motivating transformative actions towards harmful technologies, toolkits in HCI tend to perpetuate the problems that technologies create [95, 97, 119]. As [79] explains, toolkits are often designed to act as a fundamental building block of innovation; as such, they seek to accompany and promote technology adoption rather than to question or transform it [95]. For example, [119]'s analysis reveals that, in the context of AI ethics most toolkits aim at guiding practitioners in how to minimize ethical problems and provide technical fixes to them, rather than providing scaffolding for unpacking the social aspect of ethics. Further, in working as a manual, toolkits determine how practitioners see the ethical terrain and what matters to them about it, often leaving important ethical issues behind [81, 119]. [119] argues that, by decontextualizing complex social issues, these toolkits act as a "technology of de-politicization" that reduce problems to guidelines without allowing users to realize and question the power dynamics shaping their technology use.

For [95, 119], the problem lies in believing that toolkits are enough to address complex technology issues. This view leads to a technosolutionist, assimilatory tech usage only[119]. As [95] argues, the way moving forward is to "lean further into the mess", disrupting the idea of good designs, integration into existing practices, and seamless systems and rather exploring how to redistribute knowledge power. Further, it becomes critical to also foster a strong social infrastructure that can act as a support for navigating the messiness [103]. Our research explores how a toolkit can engage non-profit practitioners in some level of messiness when exploring social media data analysis as a technological innovation.

3 Developing The Toolkit

The design of the Bitacora toolkit builds off four years of qualitative research. This research began with humanitarian activists in Mexico who were increasingly using social media platforms to exchange information and self-organize to respond to local crises [2, 3, 5] . Later, the research expanded to include the Accelerator Lab Mexico, which was using social media for communication and online presence only but was interested in learning how to leverage social media data to inform their work [4]

In connection with members of the Accelerator Lab Mexico, we developed and tested a methodology to *mobilize* the knowledge produced on Twitter into the organization to inform project planning [4]. The organization chose Twitter as a platform of interest considering Facebook's policies preventing data collection and Twitter's important presence in the country [6]. This work concluded that to support responsible social media data adoption, data experts need to engage non-profit practitioners in an in-depth, critical reflection on the limitations of this type of data. As the first author reached the end of her collaboration with the Accelerator Lab Mexico, she sought to leave the organization a tool for continued reflection on responsible social media data adoption without the presence of a data expert. Drawing on the work of [103], she decided to create a toolkit as a way to provide continuity when leaving the field. While toolkits have been critiqued as artifacts to promote transformation [95, 97, 119] that are not sufficient for addressing complex social issues around technology [119], they can be useful to record lessons already learned during a previous intervention [103]. This led to the design of the Bitacora toolkit which translated the lessons from the fieldwork into supporting infrastructure for practitioners to critically and responsibly approach data from Twitter.

In what follows, we describe the main goals guiding the design of the toolkit. We also include a detailed description of the toolkit's components. Finally, we detail the data collection and analysis methodology that the toolkit proposes to non-profit organization practitioners interested in using Twitter data as actionable evidence for planning on-the-ground efforts.

3.1 Bitacora's Main Goals

The design of the Bitacora toolkit intended to meet the following objectives:

(1) *Help practitioners understand the potential and drawbacks of social media data in the context of non-profits:* As extensive previous research on adopting user-generated content as

a data source has stated [61, 112], understanding the challenges of accuracy and representativeness is critical to help practitioners set realistic expectations of what type of problems can or cannot be examined with Twitter data. Since user-generated content is unstructured and difficult to verify, it is complicated to ensure the trustworthiness, and thus, accuracy of the data : issues of misinformation and disinformation can be prevalent across this type of data [43]. Ensuring representativeness is another challenge due to the sample bias of the users and content that becomes visible or not, depending on the tools and platforms used to collect data.

(2) *Support practitioners in determining whether Twitter data is appropriate for their goals:* Non-profits are particularly vulnerable to adopting new technologies that do not fit their human connection values and practices [49]. As such, it is critical to provide practitioners with appropriate tools and guidance so that they can go through different examination phases to gradually and independently decide whether or not innovations such as Twitter data are what they need to inform their work. Design-wise, thus, the goal is to add friction and slow down the adoption impulse, encouraging practitioners to challenge the notion that this data can address all kinds of questions in the humanitarian realm.

(3) *Promote a situated analysis of Twitter data:* The findings from our fieldwork demonstrate that when analyzing data from social media platforms, it is crucial to acknowledge the interplay between online and offline interactions of the communities producing user-generated content when *mobilizing* social media data from its site of production to the site of use [2, 4]. That is, to consider the social, historical, and cultural conditions that motivate people to turn to social media platforms to organize collectively in the face of a crisis [30, 39, 69]. Engaging with such complexity and situatedness of data is critical for responsibly mobilizing data from online platforms into institutional contexts such as non-profits. Moreover, making the findings of a computational analysis usable and actionable requires transitioning them into the larger ecosystem of norms and practices of multiple actors, infrastructures, databases, legal and policy frameworks, and practitioners' operational definitions of evidence [2, 4]. As such, the toolkit sought to guide practitioners in examining Twitter data from a situated perspective, which entails considering the context where the data has been produced, reflecting on practitioners' definitions of evidence, and what they intend to do with the insights from the analysis.

3.2 Bitacora's Components

The toolkit consists of two components: a manual and two online computational tools hosted on a website [2] See Figure 1 in Appendix A for a more detailed overview of the toolkit components. The manual provided guidance and supporting worksheets for practitioners to follow and document the proposed data collection and analysis methodology. The computational tools enabled easy, automatized

---

[2]We refer to these online computational tools as Module 1 and Module 2.

access to searching, collecting, and analyzing Twitter data. See Figure 3 and 4 in Appendix A.

3.2.1 Online Toolkit Manual To help practitioners in conducting the proposed methodology, this component entailed a PDF that included an introduction to the manual, a thorough explanation of the toolkit's methodology, and a set of tools for guiding and documenting the proposed methodology. Next, a detailed description of each one of the manual's sections[3]:

(1) *An introduction to the manual* with a description of the tools included, and an explanation of the target audience of the toolkit.

(2) An *introduction to social media data* with four use cases shedding light on four scenarios in which user-generated content from Twitter and Facebook has been used to inform large and medium organizations. Among the scenarios were using Twitter data to understand citizen strategies to navigate the COVID-19 crisis in Mexico city and using Facebook data to monitor human rights violations, amongst others. Each scenario emphasized the strengths, limitations, and potential risks that come with the adoption of social media data in the context of non-profits (e.g., the incomplete and possibly misinformed nature of social media data in relation to citizen activities on-the-ground and the need to define clear-cut criteria for identifying valid evidence of human rights violations amongst the deluge of data in social media). In addition, this section included a *decision tree* to support users in discarding the scenarios and problems that were not appropriate to examine using data from Twitter. For example, the tree discouraged the exploration of systemic social problems such as poverty, corruption, or unemployment and encouraged analyzing problems such as ongoing local crises that affect the bulk of the population. To provide guidance on appropriate problems for Twitter data, the tree posed questions about the future use of the data and the type of problems they want to understand. [4].

(3) A step-by-step description of the proposed *data collection and analysis methodology* which consists of three steps that blend in an exploratory and a qualitative in-depth analysis enabled by computational analysis techniques. We describe this methodology in detail in section 3.3.

(4) *Qualitative Analysis Tools* including steps for conducting an affinity diagram and five worksheets (W1 to W5) to support exploration and documentation along the proposed methodology *The five worksheets (W1-W5)* were intended to be used along the steps of Bitacora's proposed methodology to document practitioners' decisions and observations and elicit reflections on the usefulness of Twitter data. These worksheets were: W1 for understanding and defining the problem to explore, W2 for defining keywords and time frames to limit data searches, W3 for documenting insights about different searches' results, W4 for carefully analyzing situated

aspects of the selected, final dataset, and W5 for reflecting if the aspects of the dataset suggest that Twitter data can be useful, trustworthy evidence to characterize the defined problem [5].

3.2.2 Online Computational Tools A web application enabled non-profit practitioners to use two computational tools packed in two different modules. In combination, both modules aimed to ease the process of filtering tweets to identify content similar to the problem users are examining.

(1) Module 1 offered an interface for users to access, search, and download Twitter data from user-defined queries. See Figure 3 in Appendix A. Access and search methods were based on the Twitter API v2 [113]. Twitter API credentials allowed users to retrieve and analyze Twitter data and offered different types of access for developers and academic researchers [6]. The module allowed users to download the A.CSV file with the search results.

(2) Module 2 took the A.CSV file generated by Module 1 and another B.CSV file with practitioner-provided *ground truth* training data. The module then uses Natural Language Processing (NLP) to process Twitter data to find sentences that are semantically similar to the ones identified in Module 1 and generated file C.CSV. In doing so, it narrows the number of tweets for non-profit practitioners to analyze. We implemented NLP for Spanish Language based on word embeddings using the word2vec algorithm [88, 89] with the Spanish Billion Word Corpus and Embeddings (SBWCE) pretrained model [23] and the Gensim library [99]. See Figure 4 in Appendix A.

3.3 Bitacora's Proposed Methodology

To pursue the toolkit's goals, the first author proposed to non-profit practitioners a highly iterative data collection and analysis methodology consisting of three steps that blended qualitative and computational analysis for examining and interpreting Twitter data from a contextualized perspective. See Figure 2 in Appendix A for a detailed description of the methodology. The two first steps of this methodology guided non-profit practitioners through an exploration of the problem and of Twitter as an effective source of valuable evidence for their research goals. The last step helped practitioners run an in-depth, situated Twitter data analysis via qualitative and machine learning techniques. Along each step, practitioners were asked to use worksheets W1-W5 to document their process and share reflections on their understanding of Twitter data potential and limitations for the problem they sought to characterize. As the visual representation of the methodology in image 3.3 shows, the worksheets served as breakpoints for practitioners to question their goals, revise their progress, and decide to continue their exploration, go back to previous steps or quit the exploration of Twitter as a valuable option for their problem at hand.

(1) *Step 1: Framing the Problem and Collecting a First Dataset.* This initial step requested practitioners to *first* use W1 and a decision tree for responsibly defining the problem they

---

[3]We include the original version of the toolkit's manual in Spanish and a translated version in English in the supplementary material. To facilitate the understanding of the most relevant elements of the manual, we have also added the description and graphics of some key elements of the manual in the Appendices C and D.
[4]We included the decision tree translated into English in Appendix C.

---

[5]We included the five worksheets translated into English in Appendix D.
[6]The toolkit was implemented using the academic research credentials, which allowed retrieving up to 10 million tweets per month and 100 requests per 15 minutes.

wanted to examine using Twitter data and *then* to use W2, W3, and Module 1 for conducting iterative Twitter search cycles to find the best keywords and time frames for collecting Twitter data about the problem at hand. Specifically, W1 recommended practitioners to define the problem by first going outside of Twitter to research various aspects of the problem (e.g., the locations, time frames where it took place, and the stakeholders involved), identifying existing evidence and other reliable data sources, and reflecting on their motivations for using Twitter data. To then support practitioners in defining if Twitter was an appropriate data source for their defined problem, this step also asked practitioners to use a decision tree that listed possible motivations for using Twitter data and recommended actionable pathways per motivation (e.g., if the motivation was to quantify cultural characteristics then the tree recommended against Twitter data but if it were to gather narratives of popular discourse about an event for exploring the perspective of a particular group, then the tree suggested Twitter data as a potentially useful option). Once practitioners had a problem definition and were aware of Twitter data limitations in regards to their selected problem, this step finally asked them to engage in an iterative search cycle that helped them see how Twitter was discussing the problem and potentially identifying issues such as misinformation or disinformation around the problem, the need to narrow the problem a bit more, and a mismatch between the discourse available on Twitter and their research goals. This entailed *(1)* using W2 to narrow down the problem into keywords and a search time frame that considered the problem context defined in W1 (e.g., involved organizations, communities, and Twitter accounts) ; *(2)* using these filters as input for conducting a search via Module 1 that resulted in file A.CSV; *(3)* inspecting the results to identify and document novel information (e.g., new places and time frames where the problem takes place, new involved Twitter accounts, and additional associated hashtags) using W3; and *(4)* repeating the search cycle if practitioners felt different filters were needed.

(2) *Step 2: Defining and Documenting Ground Truth.* Once practitioners were satisfied with their understanding of Twitter data based on the various searches they conducted in Step 1, the next step was to define ground truth. That is, to generate a file (B.CSV) with labels and categories representing the tweets from previous searches that were emblematic or encoded the dimensions of the topic investigated. To complete this step practitioners were first asked to use W4 for categorizing the observations about actors, communities, and initiatives they had already documented in W3 and, from there, to derive a characterization of how their problem was discussed on Twitter. Then, practitioners were asked to use W5 to reflect on how the categories and characterization in W4 related—or not—to their existing knowledge and assumptions of the type of evidence that their problem would require. As a result of this reflection, it was expected that practitioners: *(1)* determined whether Twitter offered valuable evidence for their objectives ; and *(2)* populated file B.CSV with the labels and categories representing the

types of tweets on Twitter that often discussed the selected problem (ground truth).

(3) *Step 3: Conducting an in-depth qualitative analysis on a subset of Tweets.* The toolkit then proposed an in-depth qualitative Twitter data analysis first using Module 2 and then worksheets W3 and W4 . Finally, practitioners were encouraged to follow the instructions for affinity-diagramming located in the Manual component. To use Module 2, practitioners needed both A.CSV (the initial dataset generated in Step 1) and B.CSV (the ground truth file created in Step 2) as the input. The result was file C.CSV, which stored a subset of tweets with content suitable to explore the problem previously defined. Then, the toolkit recommended to qualitatively analyze the resulting tweets via two activities. First, by reading the tweets in C.CSV one by one and using W3 once again to abstract and document novel insights (e.g., new locations, new actors, new hashtags) and W4 to categorize these insights and iterate on a possible Twitter-based characterization of the problem. Second, by using the insights from that initial analysis to organize the tweets in an affinity diagram. For this last activity, practitioners were provided with the manual's instructions on how to affinity-diagram a dataset of tweets. These entail directions for registering, organizing, and categorizing observations by theme, and then for analyzing and synthesizing the patterns and trends in each theme. The resulting synthesis shed light on different, critical contextualized aspects of the problem (e.g., the relationship between existing initiatives and the strengths of the communities). Selecting tweets for qualitative analysis is an iterative process, thus, it requires a manageable number of tweets in C.CSV so that practitioners can analyze them in detail. If the number of tweets resulting from Module 2 was very high, practitioners were recommended to further reduce the number of tweets with additional filtering (e.g., choosing a random number of tweets of the subset or only tweets that included a word or term that referred to representative locations, stakeholders, or hashtags).

4 Toolkit Evaluation

The first author conducted *two* evaluation studies with three organizations in February and March of 2022 (ten practitioners in total) to understand how practitioners from non-profit organizations with different degrees of exposure to social media and social media data analysis reacted to the toolkit goals and used the components oriented towards motivating critical reflection. The first study was conducted exclusively with practitioners from the Accelerator Lab - Mexico, with which the first author interned for over nine months. The second study was conducted with practitioners from two non-profit organizations in Mexico: *SocialTic* and *CIEP*.

(1) *Accelerator Lab - Mexico* which belongs to the Accelerator Lab Network of the United Nations Development Programme (UNDP). The Lab Network consists of 91 labs that support 115 countries and it is the largest learning network focusing on sustainable development challenges [1]. Similar to other

organizations [19], it used social media for having an online presence and communicating with other stakeholders.

(2) *CIEP* is a non-profit research center that provides information and analysis to influence, improve, and democratize discussions and decision-making in the economy and public finances sector. The organization focuses on diverse areas of public finance, such as public debt, public spending, and income and taxes. This non-profit had no experience using social media, not even for ensuring an online presence.

(3) *SocialTic* is a non-profit organization that aims to empower grassroots organizations in Latin America by reinforcing their analysis, social communication, and advocacy actions through the strategic use of digital technologies and data. Their focus is to train and accompany groups and individuals in *info activism*, the use and opening of data, and digital security. Given such a focus, this non-profit had ample experience conducting a manual, qualitative exploration of social media to help other organizations learn about potential problems for communities.

Both studies sought to explore how the toolkit impacted how practitioners from organizations with diverse social media practices perceived Twitter data's potential and limitations to be an appropriate source of information for humanitarian issues. Table 1 in Appendix B, describes each practitioner's role, background, and problem exploration methods of expertise. Next, an overview of each participant organization and their existing social media practices.

The first author moderated the evaluation sessions, which took place online via Zoom and were video recorded. To document the activities, participants used MURAL, a collaborative digital board. Each participant had a designated workspace with the five worksheets of the manual to help them complete the evaluation activities and document their experience using the toolkit. Before the first workshop session, participants received an email with the digital version of the manual, the link to MURAL, and their credentials to access the toolkit. To avoid any potential misuse of the toolkit, participants' access to the platform was granted only for the evaluation days, and the number of tweets downloaded was restricted. Participants were also continuously reminded not to share their credentials with anyone else.

Both evaluation studies started with one introductory session where participants learned about the proposed methodology, the implications of using social media data, and the evaluation timeline and goals. Individual and group sessions followed the introductory session. During individual sessions, participants used the toolkit independently, exploring the implications of integrating Twitter data into their work without mediators. In Group sessions, participants reported on their experience using the toolkit, presented their preliminary findings, asked questions, and listened to each other's progress. Introductory and group sessions lasted between an sixty and ninety minutes.

4.1 First Study (with the Accelerator Lab in Mexico)

The first assessment of the toolkit was conducted with five members of the Accelerator Lab and lasted one week. After the introductory

session, participants engaged in independent work for four days. Following the four-day work period, the first author moderated a group session where participants reported their findings and challenges working with the toolkit. The first author then conducted semi-structured interviews with participants to learn about their experiences and expectations with the toolkit. Interviews lasted between 45 and 60 minutes, were conducted in Spanish, recorded, and transcribed.

4.2 Second Study (with the CIEP and SocialTic organizations)

The second assessment of the toolkit was conducted with five participants from the CIEP and SocialTic non-profit organizations and lasted two weeks. The motivation for the second study was to understand the transferability of the toolkit from the Accelerator Lab in Mexico, where it was developed, into different kinds of non-profit organizations. The evaluation comprised an introductory session, two individual, and two group sessions. After the introductory session, participants worked independently with the platform for ten days and then met for the first group session. After the group session, participants spent time individually with the tool once again and finally gathered for a second group session, in which they debriefed their experience using the toolkit and discussed the potential steps for them to capitalize on the findings found on Twitter. Having two group sessions instead of one was done based on feedback received from the interviews in the first evaluation. Participants reported needing more time to explore the toolkit and figure out how to integrate their findings into their work.

4.3 Data Analysis

The interviews, the introductory and group sessions of both evaluations were video recorded and transcribed, adding twelve hours of data. During the design and implementation of the evaluative sessions, the first author documented her observations through memoing. Additionally, the notes, comments, and observations that participants recorded in their workspaces in MURAL and worksheets were analyzed to inform the interviews and the overall findings of the evaluation. The names of the participants were anonymized to protect their identities, and the quotes included here were translated into English. We substituted the names of participants for IDs, which are listed in the first column of the Table 1. The second column of the Table 1 refers to the organization they belong to. For clarity, as we report on participants' testimonies, we refer to them with their ID and organization name.

5 Findings

Our findings highlight practitioners' tensions when following the toolkit's guidelines and using the toolkit's materials along the three steps of the proposed methodology. When defining a problem suitable for Twitter data (Step 1), practitioners harnessed the iterative nature of the toolkit to navigate the request for using keywords and time frames, which they thought were too limiting. When choosing a ground truth that could adequately shape Twitter search (Step 2), practitioners tended to adjust their goals and problems to Twitter data's limited capacities with little reflection on the trade-offs of their decision or discussions around these data's alignment with

their goals. Finally, when facing the toolkit's request to derive insights from Twitter data using a qualitative data analysis (Step 3), practitioners saw the benefit that this request had for eliciting reflection but concluded its use in the long-term as unfeasible: they needed a process that took less time investment and that could produce more quantifiable results to legitimize future decisions.

5.1 Reactions to Step 1: Questioning Data Collection Restrictions

The first step of the toolkit's methodology asked users to define a problem they would like to examine with Twitter data. After checking if their problem definition was suitable for a Twitter analysis using W1 and a decision tree, practitioners needed to narrow the problem down to a set of keywords and a time frame for collecting the related data using Module 1, which operates in alignment with the Twitter API functionality. However, our data analysis indicates that, against the guidance of the decision tree, practitioners tended to explore quite broad and complex social problems. As such, dissecting and discretizing these types of problems and transforming them into pragmatic keywords and time frames turned out to be highly difficult for them. Specifically, breaking their problems into keywords required them to first learn with more depth *how* and *how often* people in the platform discussed these topics. Our data analysis suggests, however, that practitioners' realization of the misalignment between their expectations and the proposed methodology—which was based on Twitter's search options, drove them to suggest and explore more iterative and reflection-driven methodologies. Even though their explorations did not necessarily produce exact answers, these allowed practitioners to explore the problem on their own terms.

5.1.1 Adding Complexity to Keywords As Table 2 in Appendix B shows (second column), participants were interested in examining a wide range of problems. Practitioners interested in broad and complex problems, such as sustainability and development issues, struggled when asked to describe these problems in worksheet W2 using just keywords. As P3 explained, a critical barrier was not really knowing the terms that people on Twitter might be using to talk about these issues mainly because, despite always being present in our societies, development problems are not topics people frequently discuss on Twitter using practitioners' terminology.

> "I don't know if there's like a conversation about this, because honestly what we do in the organization, I mean these are not topics of conversation on social media." - P3, Mexico Accelerator Lab

Participants P7 and P8 echoed P3's sentiment for they also struggled to identify the appropriate terms to conduct the search.

> "I don't know how to search for some of these topics. I mean some of them are universal concepts of economics, but then, for other topics in which we are interested, the terms have multiple meanings, for example, sustainability."- P7, CIEP organization

Others, such as P8, initially assumed that people on Twitter were not discussing the topics his organization was focusing on.

> "Yes, well, it is still a very small group of people who talk about these issues, and most of them are from

academia or organizations similar to ours." - P8, CIEP organization

As participants reflected on how to fill out W2's keywords requirement, they began to discuss other pathways for using keywords that could be more cognizant of the complexity of social problems. P4, for example, recognized that it is critical for practitioners to acknowledge the various forms in which different communities talk about social issues. She, thus, suggested engaging more deeply with the toolkit's recommendation for practitioners to continuously revise their process and repeat steps if needed. Specifically, she suggested to first iteratively map the specific terms that different communities use instead of using the keywords that they, as practitioners, knew.

> "For example, when we were doing the Twitter analysis for the citizen initiatives during COVID-19, we found that when people talk about it, they use words like crisis, pandemic, virus, etc. So, it helps a lot to do several searches and map how different communities may be referring to the same topic so we don't fall into the trap of only finding about those who talk about the topic as we talk about it or think about it."- P4, Mexico Accelerator Lab

As her reflection highlights, when and if using keywords during social media data analysis, it becomes critical for practitioners to pause to learn more about the used keywords. In doing so, they can better understand who uses some keywords and who does not so as to recognize what complexities are left behind from the very beginning.

5.1.2 Shifting the Problem and Redefining the Outcome In addition to keywords, the toolkit required practitioners to use W2 to define a time frame to filter the search of the tweets. This is so as to be able to center the problem exploration around the wave of tweets arising when a crisis happens; the frequency with which topics are discussed on Twitter varies drastically: people tend to discuss mostly what concerns them at the time and trying to find comments on issues of the past in the current conversation might not be effective.

However, for participants such as P1, the time frame, and keywords restriction, together with the toolkit's emphasis on continuous revisions of progress, were determinant in how they engaged with the proposed problem definition and data collection methodology. Specifically, these restrictions drove him to shift problem definitions and time frames several times while defining an entirely new methodology in the process. P1 initially sought to examine the intersection between gender and resilience in general. However, after using W1 and the decision tree, he realized that these topics were too abstract and needed to be refined into a more concrete problem. He then decided to examine the role of gender inequality during crises and was particularly interested in the recent COVID-19 pandemic crises given that, at the time, Mexico was still experiencing the aftermath of this event.

In reflecting on the time frame, he concluded that using the pandemic as a reference within the current time frame was not going to be effective. As he explained, "[..] on Twitter, people tend to post what they have at the top of their minds, and I thought it

would be difficult to find something interesting [for the COVID-19 crisis was no longer recent]." The new problem definition was also not really helping him to find the needed keywords: "It is difficult to capture the conversation of how the impact on women and men during a crisis is uneven." Thus, before defining keywords or choosing a time frame, P1 felt he still needed to know more about gender and inequality in the context of crises and, drawing on worksheet W1's prompts for exploring the problem outside of Twitter, he engaged in an online research. This process finally shed light on a time frame of interest for him:

> "Something I didn't expect was that a critical moment in the gender and resilience conversation happened two years after the earthquake, in 2019, when the university started publishing a series of research studies that provided more evidence of the unequal impact of the earthquake between women and men. In parallel to the publication of the studies, there was a wave of tweets and newspapers promoting this research and discussing gender inequalities. So, the fact that this discussion happened two years after the earthquake was very interesting to me."- P1, Mexico Accelerator Lab.

Building on these findings, P1 shifted focus for a second time and decided to examine gender and resilience in the context of the 7.1-magnitude earthquake that shook Mexico City in 2017. He felt that this shift would be useful also for his initial goal of understanding the role of gender in the COVID-19 crisis: "Examining the conditions of inequality during the earthquake might help us to understand what we are seeing now [in the aftermath of COVID-19]."

In continuously challenging and even moving away from some of the toolkit's methodological recommendations (e.g., narrowing down problems to keywords and time frames), P1's case illustrates the critical research and contextualization steps that practitioners need to go through before narrowing down a problem and using Twitter to analyze it. As P1 reflects, when approaching a social problem, it is key to first learn as much about it as possible:

> "One of the things that I noticed with the rest of the participants is that they did not have many results because their problem was already very defined [...] It is not that you have to define too many variables from the beginning. It is more like a problem is pointing in a direction, so I am going to try all these combinations." - P1, Mexico Accelerator Lab.

Practitioners' reactions to the request of using keywords and time frames to define a problem of interest, illuminates this process is far from seamless. It needs time and resources for engaging in rich iterations, time frame relocations, comparisons, and the finding of outside sources, amongst others. Only by slowing down the need to reduce problems to search queries, it becomes feasible for practitioners to become fully aware of how discourses evolve on Twitter when a crisis arises, how institutions and citizens interact with these discourses, and connect these discourses with their issues of interest.

5.2 Reactions to Step 2: Defining Ground Truth Despite the Risk of Incompleteness

After generating a first dataset using keywords and time frames, the toolkit asked practitioners to collect tweets of what practitioners considered to be adequate to represent their problem's ground truth . This implied reading the initial dataset's tweets one by one and analyzing them using W4 and W5 to reach a definition of what counted as evidence from Twitter data concerning a particular problem. The request for practitioners to engage in a careful examination of the dataset sought to help them realize that Twitter data is never complete and that these data analyses can easily dismiss non-traditional sources of information. In particular, the toolkit proposed the ground truth definition phase as the last opportunity for practitioners to verify Twitter data's suitability for exploring problems of their interest before engaging in a more in-depth data analysis. We found however, that, at this phase, it became harder for practitioners to dismiss Twitter data as a source of interest or to engage in an in-depth analysis of the perspectives left behind when choosing ground truth.

5.2.1 Settling for Twitter Data's Incompleteness by Considering Twitter as 'Just' a Starting Point The ground truth definition step sought to help practitioners to analyze biases within their definitions. That is, to realize how the selection of one form of evidence is leaving other forms behind, and to recognize that the results will, therefore, never be complete. Across Step 1 and 2, the toolkit provided guidance to practitioners about how to recognize and use the realization of such incompleteness as a criterion to decide whether to continue Twitter data use or not. For example, the Manual's intro to social media explained incompleteness and its impact in data analysis, and the decision tree in Step 1 recommended avoiding Twitter data when the insights it could provide were incomplete in relation to the explored problem. Further, worksheets W1 in Step 1 and W3 and W4 in Step 2 gradually motivated practitioners to think about what could be considered as evidence for the problems of their interest and if Twitter data analysis could provide the defined evidence.

In some cases, such as P2's, this guidance was effective for eliciting reflection. P2's goal was to use Twitter data to build an inventory of practices and organizations involved in the recovery of communities during natural disasters. Thus, at the beginning of Step 2, he sought to prioritize the repetition of keywords and hashtags used by non-official media Twitter accounts as evidence of certain practices' and organizations' prominence.

> "For me it was the repetition of words rather than the hashtags what counted as evidence that Twitter could provide. I focused a lot on them and used them in my analysis as a criterion for elimination. If a hashtag was repeated, I included the tweet, otherwise, I discarded it. Another thing that counted as evidence was the tweets that repeated names of places, names of journalists, politicians, etc."- P2, Mexico Accelerator Lab

However, as he engaged with worksheets W4 and W5 (which asked practitioners to reflect on the type of insights they could obtain from Twitter vs their expectations), P2 began expressing

concerns about limiting the representativeness of the data. In particular, it worried him to leave behind relevant data and the way such a dismissal could prevent a richer understanding of organizations' behaviors.

> "You don't know if those things that show up in the dataset indicate that there is a pattern or something necessarily relevant for the people who are affected by the crisis, there could be other reasons that explain why certain topics are discussed more than others."- P2, Mexico Accelerator Lab

P2's reflections on data incompleteness were exactly the type of reaction that the toolkit's design sought to elicit; such a reaction can help practitioners even manage issues of misinformation and disinformation in the future. However, our data analysis illuminates that, most practitioners reacted rather differently. Due to a prevalent view of Twitter as a starting point for problem exploration, they did not see data incompleteness as a reason to dismiss Twitter data analysis. As P3 and P1 explain, practitioners do not see Twitter as the only source of exploration. Rather, they see it as *"a way to do a quick scan"* (P3) and a source for *"small signs or glimpses of much more complex stories"* (P3) , and, thus, settle for its incompleteness in the name of the information it can provide, such as "reference points in a map" (P1). P3 explains further:

> "The toolkit would be useful for having a broader overview of the problem. Obviously, we cannot interpret everything we find on social networks as the complete reality because the content there is very biased." -P3, Mexico Accelerator Lab

5.2.2 Adapting Problems and Ground Truths to Twitter Data Limitations Another practitioner's reaction to the ground truth definition stage that our analysis illuminates is that of adapting to Twitter data limitations instead of necessarily working against them. As practitioners spent time understanding how people on Twitter discussed topics, they become more informed about the type of ground truths that were possible to use, and how this availability aligned or not with their initial goals. For some participants, this helped reinforce some of their perspectives and biases rather than helping them to reflect on the potentially negative consequences of their decisions.

For example, P3's work during Step 1 suggested she already had a bias against unofficial information as they could more easily spread out rumors and misinformation. After looking at her dataset during Step 2 and noticing the difference between tweets coming from official and non-official sources, she confirmed her bias: Twitter's emphasis on translating popularity into value and validity pushed her to dismiss unofficial sources as relevant to her process.

> "Some times a topic for some reason can suddenly become a Twitter phenomenon. For example, there may be a tweet with thousands of likes and, on the other hand, a note that talks about the same topic but with very few likes. But the note may have more relevant information, even if it has not become viral."- P3, Mexico Accelerator Lab

As a result, and despite the reflection-eliciting questions for W4 and W5 guiding her to account for all interesting sources and insights, she decided to only consider content coming from official sources as valid evidence. Specifically, she filtered out any content that was not posted by verified accounts (with a black check) of government entities, well-established non-profit organizations, or recognized media sources and content posted by non-verified accounts that posted references verifiable beyond Twitter (e.g., traceable content to other sources such as statistics, government and other organizations reports, etc.). For her, thus, evidence of public discourse became less valid. In avoiding the risk of misinformation in Twitter, however, she was (possibly unintentionally) also dismissing important thoughts, reflections and experiences that, while not official or formal, could be of relevance to inform problem exploration.

We also found that, at this stage of the methodology (Step 2) practitioners were more willing to adapt or align their goals and process to the type of ground truth available on Twitter than to dismiss Twitter data analysis as a whole. For example, P5 reported the following after having done a couple of searches and analysis of Twitter data.

> "In the beginning, before reading the manual, in the first worksheet [W1], I wrote in the description and expectation section of the project something like identifying the root of the conflicts between indigenous populations and private initiatives. I think my goal was a bit far from reality. After reading the manual [which included an introduction to Twitter data and its limitations], I changed my objective and wrote that I wanted to identify the public discourse of the private sector and the government regarding indigenous consultation."- P5, Mexico Accelerator Lab

As her account highlights, it was her initial understanding of the type of evidence that Twitter could provide what drove her to narrow down her research question a bit more. This question never changed, not even after the ground truth definition in Step 2, which via W4 and W5 asked her to reflect on how the experiences in the data could serve as evidence for their research questions. In focusing her work on her narrowed-down question she might have dismissed the important process of learning more about the problem.

Worksheets W4 and W5 sought to support practitioners in using the ground truth definition phase as a moment to slow down, reflect on the different, novel insights that Twitter could provide, and analyze the alignment with the type of evidence they needed to explore the problems of their interest. However, at this stage, this was not that feasible for them: their view of Twitter as "just as starting point" and the little impact that materials such as W4 and W5 had on helping participants change their already defined biases or decisions at such an advanced point in the methodology, drove many practitioners to settle for Twitter data's limitations and often adapt to them rather than rethink their whole goals and processes.

5.3 Reactions to Step 3: Rejecting a Situated Data Analysis due to Uncertainty and Exhaustion

The last step in the toolkit's methodology is an in-depth situated qualitative analysis. This entails reflecting on the social, cultural, and political conditions of the communities that produced the data [73]. To this end, the toolkit first asked practitioners to use worksheets W3 and W4 to derive new insights from the data and finally, to follow instructions in the Manual to put together an affinity diagram for identifying patterns and themes from the set of tweets generated by Module 2. While this step indeed offered participants an important opportunity to really discuss the associations and implications of using the collected data, for the most part, participants reported uncertainty, frustration, and exhaustion along the process. Despite these issues, engaging in qualitative analysis did force practitioners to reflect on the tensions between using data for generating clear-cut decisions versus appreciating and respecting data's complexity.

5.3.1 Facing Subjectivity with Uncertainty All of the participants involved in the toolkit assessment referred to the qualitative analysis as too demanding both in terms of time and effort. P9 explains this problem further:

> "From what I understood using Module 2, the analysis consists of reading everything and extracting relevant phrases. But that's also a lot of work, and if it is only one person doing it, it's too much. I mean, you're not going to analyze one thousand tweets by yourself." - P9, SocialTic organization

Although the toolkit provides tools to guide and support users in this step (worksheets W3 and W4 to help identify and summarize insights from a dataset and the manual's instructions to help in conducting an affinity diagram of tweets), it also assumes that regardless of their training or background, practitioners will be able to analyze and synthesize insights from Twitter data. That was not the case. While five out of ten participants had experience with qualitative methods, and three had analyzed Twitter data, all reported struggling to decide what insights to pursue as the qualitative analysis progressed. Not having a way to identify or quantify the most repeated words, hashtags, or Twitter accounts in the dataset they were analyzing, made it difficult for practitioners to decide what to care for or discard (e.g., highly repeated topics that were based on rumors).

> "When I was analyzing the data, I kept wondering about the accuracy of the reality that a set of tweets might show us; I mean without having a number it is hard to assess the relevance of an observation. For example, I paid attention to the tweets that came from a certain group of people, but I paid attention to the content and not the number of tweets. So, I don't know if the conclusions I'm making are relevant or not." - P7, CIEP organization

Participants requested tools such as histograms, word clouds, or graphs to reduce human labor and speed up the decision on which topics to analyze further. P9's account sheds light on practitioners'

need to quantify topics to legitimize their decisions, making them less subjective:

> "An idea I had is that the toolkit could have a tool that divides the tweets into tokens, bigrams, or something to know what phrases appear in the dataset. That could help us get a sense of what people are talking about in the data. I know this is not the toolkit's goal, but we need something to get hints of what is discussed. Otherwise, just by reading, it is possible we miss some relevant information." - P9, SocialTic organization

Quantification tools, however, might be counterproductive to the goal of eliciting in-depth reflection for a responsible use of social media data; identifying the repetition of a word, a hashtag, or a Twitter account runs the risk of being immediately interpreted as a relevant aspect in relation to the examined topic when in fact, visibility on social media platforms does not equate prominence or relevance. Internet access, skills, and demographic factors heavily influence the populations represented in each platform [48, 61].

5.3.2 Navigating Tensions between Learning from Data and Time Constraints An additional concern that participants reported was the extraordinary amount of time that conducting a qualitative analysis demanded. Even though Module 2 filtered the tweets by identifying those most similar to practitioners' topic of interest, some of the participants still considered the time invested as significant. However, our data analysis suggests that the time needed increased due to also an increase in participants' need to double check and expand on the insights from the data analysis by using other information sources outside of Twitter.

> "There is intermediate work between continuing to advance with downloads and word filtering. There is human work that we have to do between steps. For example, after downloading tweets and reading some of them, we spent time searching in the newspapers and other media about which cases [about feminicides] were the most relevant. Then, we used that information to decide what to analyze in more detail with the toolkit." - P9, SocialTic organization

While time-consuming, these intermediate steps are, in fact, part of the planned outcome of the toolkit. Worksheet W1, for example, motivated practitioners to list existing evidence and data sources, and W5 asked for a comparison of insights with existing knowledge). Constant contrast with the world outside of Twitter is needed for a critical reflection of social data so as to better define whether or not the results of Twitter analysis are valid, representative of reality, and can be actionable. The fact that participants resorted to these steps suggests that they recognized Twitter data interpretation as a complex and holistic activity (e.g., requiring information outside of Twitter).

In addition to expanding data sources, participants reported discussing their process and insights with other members of their organization. P6 explains this behavior in more detail:

> "I think something worth mentioning is that it is not only that I did the searching and analysis of tweets. At the same time, I needed to brainstorm with other

team members and ask for feedback to see whether I could use alternative keywords or frame the problem differently. I think those things are necessary to validate the work with the toolkit and require additional time. Thinking about the problem and how to search for data for it is what took us more time I think." - P6, CIEP organization

The existence of intermediate steps and in-group discussions signals that the toolkit's push for qualitative analysis is promising. However, participants also expressed fear that such time investment would still not be able to answer their specific research questions.

6 A Toolkit for Slowing Down Social Media Data Innovation: Discussion

In the last decade, there has been increasing research that praises the potential of user-generated content for informing the operation of non-profit organizations in the humanitarian sector [3, 52, 56, 114]. Despite the increased development of computational methods and techniques, there has been less discussion about what would imply for organizations to adopt these tools and methods to render social media data actionable. In particular, emergent exploratory work has stressed that non-profit organizations and practitioners need to learn how to continuously challenge three assumptions about the power of social media data for informing humanitarian action [4]: (1) that computational methods and tools are enough to extract actionable insights from social media data ; (2) that social media data is flexible enough to provide insights about any possible humanitarian problem; and (3) that social media data can provide lessons regardless of the social and cultural conditions that created it.

The Bitácora toolkit is aimed at helping practitioners to unpack and challenge these assumptions. Specifically, it sought to guide them so that they would choose to use Twitter data only after understanding these data limitations, its suitability to address the problems of their interest and the context in which these data were produced. Both explicit design decisions (e.g., motivating the use of sources outside of Twitter in W1 and W5 and recommending to contrast existing knowledge with the analysis insights in W5), and unexpected behaviors from practitioners (e.g., navigating keywords and time frame constraints and taking time to validate insights with colleagues) contributed to the success of the toolkit experience in helping practitioners face and explore some of the assumptions around social media data use. However, the amount of time, labor, and care, together with the impossibility of the toolkit to produce pragmatic, discrete results that could hold quantifiable validity, drove practitioners to reject the toolkit as a viable, long-term solution for them.

We now discuss the possibility of moving forward with Bitácora, a toolkit that seeks to slow down technological innovation at the intersection of datification, institutionality, and the need to make quick decisions affecting lives on the ground. Specifically, we reflect on the design aspects that were successful at helping practitioners to unpack the implications of social media data as a technological innovation, discuss feasible pathways for the toolkit to have an

effective impact on non-profits' operation, and suggest modifications for the next iteration of the toolkit based on the participants' feedback.

6.1 Successful Design Aspects

As [49] argued, the rhetoric of technological innovation can create many problems in the space of non-profit organizations: it pushes non-profits to forever pursue perfect but impossible futures. In the process, it creates 'new normals' that end up adding more work to their everyday practices and profoundly disregarding their goals, and values, as well as the care and trust-building efforts that foster their human connections. As such, authors are increasingly making calls for designing against the status quo [64]; that is, to no longer create technologies that fit into existing routines—which are, in fact, inequitable and unsustainable, but that support radically new ways of working. Introducing social media data analysis in the non-profit sector, as we explained, is one of such innovations that runs the risk of being highly harmful if not done with extreme care.

Through a series of explicit limitations, such as a decision tree for problem definition, worksheets for documenting decisions, and an emphasis on guiding a qualitative data analysis, the Bitácora toolkit represents a first step for supporting non-profits to disrupt the status quo. In particular, it aimed at helping non-profits challenge the rhetoric behind social media data and slow down its adoption. While the explicit limitations built into the toolkit's design did not work as expected, we discuss two design aspects of the toolkit that our findings suggest as especially helpful for motivating participants to reflect more deeply on the limitations and advantages of using social media data to address social problems. None of these design aspects, however, worked in a straightforward manner; each generated tensions that can also shut down practitioners' initiatives to go beyond the toolkit and push them away from reflecting on the use of social media data, situatedly.

6.1.1 Design Aspect 1: Fostering an iterative process while including overly-limiting requirements. As explained in Section 3 and shown in Fig 2 in Appendix A, the toolkit's methodology was highly iterative: by posing questions that required deep reflections such as "Whose perspective do I seek to understand?" in the decision tree (Step 1), or "Did you identify new actors or impacted communities? in W4 (Step 2 and Step 3), What is interesting about them?", the toolkit motivated participants to continuously revise their progress. At the same time, having a series of clear-cut, limiting elements along the process (e.g., a distinct workflow for deciding when to use or not Twitter data, or the request to narrow down problems to keywords in W2), acted as breakdowns that drove some practitioners to actively decide to question those limits and follow their own iterative process; these limitations acted as generative breakdowns. For example, when defining their problem of interest (Step 1), the keyword and time frame requirements in W2 drove practitioners to not only challenge the methodology and propose new ones, but also engage in a highly iterative problem definition process that enabled them to use Twitter data in non-deterministic ways: as a keyword provider for Google searches, an enabler of comparisons across locations, events, and time, and so on.

In regards to [81, 119]'s critique of toolkits over-limiting the scope of practitioners' understanding of a problem, our findings

suggest that a possible way to move away from limitations and elicit exploration outside the toolkit is intentionally designing breakdown moments. Our findings also suggest, however, that toolkit designers need to remain cautious about how they foster these breakdowns to be generative: if the toolkit does not give enough support to practitioners to push against the limitation, it could end up shutting down practitioner's desire to explore and iterate. Further, as seen in the data, these supports need to be stronger and more visible as practitioners progress across the methodology for it becomes harder for them to challenge social media data use once highly advanced in the process.

6.1.2 Design Aspect 2: An emphasis on a manual, tweet-by-tweet qualitative analysis while constantly motivating to contrast insights with outside sources. Based on our previous research [2, 4], one of the toolkit's central tenets was to guide a situated use of social media data. That is, one that takes place after a rich understanding of how topics are discussed in as well as outside of social media, and how this available discourse fits with the definition of evidence of the organizations that might be able to leverage such content. To support this tenet, in Steps 2 and 3, the toolkit asked practitioners to read and categorize tweets one by one and document new findings about the problem's locations, actors, and time frames (W3). Further, it included a series of questions along the provided worksheets motivating practitioners to reflect not only on the problem context, actors, and existing evidence (e.g., W1) but on how the insights of each line-by-line analysis contrasted with their current knowledge of the problem (e.g., W4 and W5) . As our findings highlight, these design decisions forced practitioners to move back and forth between the data, and other information sources, discussing their insights and discovering new problems and facts along the way. As such, it was successful in making them analyze Twitter data's limited role in shedding light on the messiness of social problems. It disrupted their expectations and drove them to question many assumptions about the feasibility of Twitter to answer the questions they needed.

Despite this success in eliciting critical perspectives, this particular step also made it clear to practitioners that the toolkit and its methodology could not be part of their everyday practices. To engage in this process, practitioners had to be willing to conduct multiple and iterative analyses, which demand time and care. Not only this process was too time-consuming and mentally demanding, but it did not offer any concrete results that could suggest certainty in taking actions on the ground. In line with [32], practitioners seek for toolkits that easily fit in their working pipeline and that are quick and easy to use.

Given that these two generative design aspects are hard for practitioners to operationalize on an everyday basis, an open question for future work then is: *where in the pipeline can a toolkit such as Bitácora, that can elicit reflection but struggles to drive actionable change when in need to make on-the-ground decisions, operate better?* In the next section, we discuss possible deployment pathways as well as specific modifications that the next iteration of the toolkit needs to consider to best fit these pathways.

6.2 Long-Term Sustainability of the Toolkit

Is it possible to slow down social media data innovation in the long run in the context of non-profits? To a certain extent, our findings align with [119]'s critique: simply using a toolkit to address complex issues is not enough. Practitioners could not see how the toolkit's slowing-down efforts could fit at the intersection of datafication, institutionality, and the need to make quick, effective decisions affecting the use of human and financial resources as well as lives on the ground. Engaging with data from a critical perspective takes time, deliberation, and willingness to accept that pursuing the innovative path might not be the best way to go. The power dynamics shaping the context and work of non-profit organizations do not allow for such an inversion of time, especially considering that the result does not align with existing expectations of productivity and efficiency [49, 58].

However, based on our findings, we argue that to have a long-term impact on slowing down innovation, the toolkit should intervene at a different point in non-profit practice. Rather than aiming at operating on its own to guide urgent decisions, the toolkit could work together with a facilitator to accompany practitioners during other, more learning-oriented, lower-stakes situations. We now discuss possible points for intervention and the various dimensions to consider at such points, including equitable non-profit and social media relations. Further, we discuss aspects to consider for the next iteration of Bitácora based on this evaluation's results.

6.2.1 A Feasible Deployment Point: Considerations As practitioners indicated, it can be difficult for a non-profit to rely on a toolkit such as Bitácora to make urgent decisions about a situation on the ground. Thus, we propose Bitácora as a toolkit that aids facilitators who work with practitioners during situations in which organizations can afford the exploration of social media data as a potential option. These could be training sessions about social media data use, which take place in a safe, educational space. They can also be done in parallel to real higher-stake on-the-ground projects that aim to understand how social media data could enrich existing insights. While using Bitácora in these spaces will not lead to radical change (e.g., these situations already operate under the assumption of social media data as an unavoidable innovation), it can instill small changes in practice in the long run.

Our findings suggest there are three critical considerations before successfully introducing Bitácora in these lower-stake spaces: the organizational requirements, the types of problems and questions worth exploring, and the relation between the non-profit and social media platforms providing the needed data.

The organizations that participated in the evaluation all had enough time and resources to afford the freedom to fail in the exploration of social media data. For all three, the use of social media data was a promising alternative but not better or more needed than others they are already using to make decisions on the ground. Based on these observations, we recommend that organizations undertaking the use of a toolkit such as Bitácora for guiding low-stake situations assess their availability of time, level of pressure to use social media data for guiding on-the-ground actions, willingness to revise or reconsider their goals, and flexibility to fail. Further, it is worth mentioning that, given the change from Twitter to X, the toolkit currently has a highly limited capacity to retrieve tweets

(1500 per month). To increase its capacity, organizations would also need to have enough financial resources to invest either $5,000 or $42,000 a month. Finally, it is critical to carefully select the practitioners who will participate in the exploratory spaces where the toolkit will be used. Our findings indicate that the toolkit can be of important support to teams with members holding varied forms of expertise (e.g., members experts in problem framing, community advocacy, and so on) and at least one member with an appreciation of qualitative insights.

The types of problems and questions that the toolkit can better support, either during training sessions or exploratory work taking place in parallel with other initiatives, are also important to consider. Given the toolkit emphasis on keywords and time frames so as to more easily find exemplar tweets, in our evaluation practitioners tended to gravitate towards systemic problems taking place in the context of crises, which can be easier to pinpoint and narrow down. When exploring these problem contexts, the nature of Twitter data helped to illuminate various, interesting problem dimensions such as people's experiences and strategies, public discourse, and actors contributing to relief efforts.

Finally, *the business model* needs to be carefully defined, especially after the dramatic transformation of Twitter to X in such a short time for it shows the dependency and constraints that non-profits are subjected to when using data from social media platforms. The evaluation of the toolkit took place during February and March of 2022. By the end of that year, Twitter was acquired by Elon Musk, and under his administration, the platform has been through a substantial transformation, not only having its name changed to X but also having many of its core features dismantled [92, 122]. One of the key changes is the discontinuation of academic access to the Twitter API [77] and the increase in costs for each type of access: with these new constraints retrieving the same number of tweets we accessed freely in the past would now cost us at least $42,000 a month [108]. The current version of the toolkit, thus, will have to go from retrieving 10 million tweets per month to 1500 tweets per month. It is, thus, critical to consider these shortcomings when encouraging non-profits to adopt these data sources; non-profits could engage in social media data use when it is still accessible and then be forced to pay once policies change. Thus, even before recommending using data from any social media platform, those motivating interventions such as a toolkit, are responsible for informing practitioners about the potential risks of using that kind of data. Knowing the potential risks associated with using social media data is key for practitioners because integrating a new methodology or type of data into their work entails an investment of their time and resources, which tend to be limited. Thus, practitioners need to know that when using social media data, there is always a risk of losing access to it if a company makes changes similar to what happened with Twitter. We recommend HCI researchers and data experts supporting non-profit organizations in their data practices to always communicate with them about the constraints, rules, and associated costs of using that data. A remaining question to reflect on is how we reduce data dependency and prevent non-profit organizations from being dependent on data that they can lose access to at any time.

6.2.2 A Next Iteration of the Toolkit: Critical Changes Building on our analysis, we provide the following considerations to inform future iterations of Bitácora. We organize our recommendations following the structure of the toolkit's methodology. It is worth mentioning that, given the change from Twitter to X, a new iteration of the toolkit will operate at a very low capacity; it will go from enabling the retrieval of 10 million tweets per month to 1500, which can be quite limiting even if used only during training sessions.

Step 1: Problem Framing. As previously discussed, constraints in Step 1 operated as generative breakdowns that motivated some practitioners to engage in deeper problem explorations. A key change would be to redesign these constraints so as to begenerative for more practitioners. In the case of the keywords constraint, this entails providing practitioners with different examples of how to interact with it based on previous practitioners' experiences. For example, the toolkit could suggest practitioners to first collect terms used frequently by people before defining keywords, or using an initial set of keywords to search for the topic outside of Twitter to identify problems that could more easily be narrowed down to keywords.

Step 2: Ground Truth Definition At this stage, participants showed a willingness to adapt their goals, problems, and definition of evidence to Twitter limitations without further reflecting on the trade-offs of such decisions.. This highlights a need for practitioners to hold a stronger grasp on the implications behind deciding what counts or not as ground truth: ground truth not only defines what realities will be represented, it determines what misinformation or disinformation will be perpetuated in the data analysis [28]. The change from Twitter to X increases the need for practitioners to better understand how to rate the trustworthiness of the multiple definitions of ground truth. The transition of the verification system into a membership model and the dismantling of teams moderating content runs a high risk of increasing the amount of misinformation, disinformation, and impersonation on the platform [13, 42]. This might reduce the representativeness of data and pose new questions regarding what algorithmic curation means under the new model of interaction. In a next iteration of the toolkit, we will include examples of other social media data analyses that were used as evidence to inform the organizations' actions and to draw insights. Further, we will add informative cases to help them delve deeper into the human role of defining ground truth and how these definitions are made in light of the arguments practitioners intend to make, the audience they might intend to inform, and the risk of misinformation and disinformation around the problem explored. Each informative case will have a reflection on what is left out when following different definitions of evidence.

Step 3: Qualitative Analysis. Practitioners reported feeling overwhelmed when asked to analyze a large number of tweets qualitatively. In response to this demand, they asked for computational tools to filter the tweets (e.g., histograms, word clouds), reducing the data they would analyze manually. Given that our commitment to developing this toolkit was to promote a situated data analysis, we consider that computational tools should be used with caution and with the guidance of a data expert. Thus, in the future, we will include additional tools to filter the data. Additionally, we will encourage practitioners not to immediately consume the outputs but

instead interrogate them and reflect on the limitations of quantitative analysis. For example, if we were to use a word cloud to identify trends, we would suggest reflecting on the underlying factors that contribute to those trends rather than merely consuming them as facts. Practitioners could ask, 'what type of accounts tweeted the most?' 'Which words are less frequent?' By interrogating the outputs of quantitative analyses, we could encourage a situated approach to social media data.

7 Limitations

This research's findings offer a first yet limited understanding of initiatives for slowing down social media data innovation in the context of non-profits. Important aspects that limit this research are the use of Twitter data, the type of organizations involved in the design and then evaluation of the platform, the online nature of the evaluation sessions, and the fact that this is a qualitative work. We selected Twitter as the platform for the toolkit's design giving its prominent use in Mexico and this platform's openness to provide services to the public at the time of the study (from February to March of 2022). As such, while this research's findings can inform interventions using other social media platforms, they cannot be generalized to all social media data. Further, changes in accessing the Twitter API [113], the verification system, and the content moderation after Twitter's change to X, greatly impact the toolkit's operations; under the new business model of X, we can no longer retrieve the same amount of data [77] and the new membership model and dismantling of teams countering disinformation runs the risk of increasing misinformation and disinformation [13, 42], requiring a new iteration of the toolkit to strengthen its emphasis on these issues. As for the type of organizations participating in this study, they hold many privileges that most non-profits do not: they have time and freedom to explore new alternatives to inform their actions and are often seeking innovative forms of doing their work. It is, thus, critical, to replicate this work with other non-profits that have more restrictions that the ones participating in this research. The online nature of the evaluation session also could have impacted the interactions and conversations amongst researchers and participants. Finally, as qualitative research, this study does not provide clear-cut comparisons amongst the toolkit's design features in regards to how they impacted practitioners' practices and beliefs. Future work could engage a critical mass of users and collect data for a quantitative assessment of the toolkit's capabilities.

8 Conclusion

In this paper, we reported on the design and evaluation of the toolkit Bitacora, addressed to practitioners working in non-profit organizations interested in integrating Twitter data into their work. The components of the toolkit support practitioners in deciding independently whether Twitter data is appropriate for their goals and emphasize a situated analysis of data. We collaborated with three organizations in Mexico to assess the toolkit's effectiveness in guiding practitioners to explore the implications of integrating Twitter data into their work without mediators. Our findings emphasize practitioners' tensions while following the toolkit's guidelines and materials along the three steps of the proposed methodology.

References

[1] [n. d.]. United Nations Development Programme. https://www.undp.org/
[2] Adriana Alvarado Garcia, Matthew J. Britton, Dhairya Manish Doshi, Munmun De Choudhury, and Christopher A. Le Dantec. 2021. Data Migrations: Exploring the Use of Social Media Data as Evidence for Human Rights Advocacy. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 268 (Jan. 2021), 25 pages. https://doi.org/10.1145/3434177
[3] Adriana Alvarado Garcia and Christopher A. Le Dantec. 2018. Quotidian Report: Grassroots Data Practices to Address Public Safety. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 17 (Nov. 2018), 18 pages. https://doi.org/10.1145/3274286
[4] Adriana Alvarado Garcia, Marisol Wong-Villacres, Milagros Miceli, Benjamín Hernández, and Christopher A Le Dantec. 2023. Mobilizing Social Media Data: Reflections of a Researcher Mediating between Data and Organization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 866, 19 pages. https://doi.org/10.1145/3544548.3580916
[5] Adriana Alvarado Garcia, Alyson L. Young, and Lynn Dombrowski. 2017. On Making Data Actionable: How Activists Use Imperfect Data to Foster Social Change for Human Rights Violations in Mexico. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 19 (Dec. 2017), 19 pages. https://doi.org/10.1145/3134654
[6] Clay Alvino. 2021. Estadísticas de la situación digital de México en el 2020-2021. https://branch.com.co/marketing-digital/estadisticas-de-la-situacion-digital-de-mexico-en-el-2020-2021/
[7] Susan Anson, Hayley Watson, Kush Wadhwa, and Karin Metz. 2017. Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors. *International Journal of Disaster Risk Reduction* 21 (2017), 131–139.
[8] Lora Aroyo and Chris Welty. 2013. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *ACM Web Science 2013*. Association for Computing Machinery, New York, NY, USA.
[9] Matthew P Aylett and Shaun Lawson. 2016. The smartphone: a Lacanian stain, a tech killer, and an embodiment of radical individualism. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 501–511.
[10] CD Balana et al. 2012. Social media: major tool in disaster response. *Inquirer Technology* 5, 3 (2012).
[11] David Bell, Mark Lycett, Alaa Marshan, and Asmat Monaghan. 2021. Exploring future challenges for big data in the humanitarian domain. *Journal of Business Research* 131 (2021), 453–468. https://doi.org/10.1016/j.jbusres.2020.09.035
[12] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[13] Thor Benson. 2023. Yoel Roth on Elon Musk gutting Twitter's election integrity team and what it means for 2024. https://www.publicnotice.co/p/yoel-roth-interview-twitter-elon-musk-elections

[14] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[15] Chris Bopp. 2019. Doing "Good" with Data? Understanding and Working Around Data Doubles in Human Services Organizations. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 33–37. https://doi.org/10.1145/3311957.3361850

[16] Chris Bopp, Ellie Harmon, and Amy Voida. 2017. Disempowered by Data: Nonprofits, Social Enterprises, and the Consequences of Data-Driven Work. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3608–3619. https://doi.org/10.1145/3025453.3025694

[17] Chris Bopp and Amy Voida. 2020. Voices of the social sector: A systematic review of stakeholder voice in HCI research with nonprofit organizations. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 2 (2020), 1–26.

[18] Kirsten E Bray, Christina Harrington, Andrea G Parker, N'Deye Diakhate, and Jennifer Roberts. 2022. Radical futures: Supporting community-led design engagements through an afrofuturist speculative design toolkit. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[19] Rowena L Briones, Beth Kuch, Brooke Fisher Liu, and Yan Jin. 2011. Keeping up with the digital age: How the American Red Cross uses social media to build relationships. *Public relations review* 37, 1 (2011), 37–43.

[20] Ryan Burns. 2018. Datafying Disaster: Institutional Framings of Data Production Following Superstorm Sandy. *Annals of the American Association of Geographers* 108, 2 (2018), 569–578. https://doi.org/10.1080/24694452.2017.1402673 arXiv:https://doi.org/10.1080/24694452.2017.1402673

[21] Ryan Burns. 2019. New Frontiers of Philanthro-capitalism: Digital Technologies and Humanitarianism. *Antipode* 51, 4 (2019), 1101–1122. https://doi.org/10.1111/anti.12534 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/anti.12534

[22] Eleanor Burt and John A Taylor. 2000. Information and communication technologies: Reshaping voluntary organizations? *Nonprofit management and Leadership* 11, 2 (2000), 131–143.

[23] Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings. https://crscardellino.github.io/SBWCE/. Accessed: 2022-04-11.

[24] Samuelle Carlson and Ben Anderson. 2007. What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication* 12, 2 (01 2007), 635–651. https://doi.org/10.1111/j.1083-6101.2007.00342.x arXiv:https://academic.oup.com/jcmc/article-pdf/12/2/635/22317230/jjcmcom0635.pdf

[25] Sarah Carnochan, Mark Samples, Michael Myers, and Michael J. Austin. 2014. Performance Measurement Challenges in Nonprofit Human Service Organizations. *Nonprofit and Voluntary Sector Quarterly* 43, 6 (2014), 1014–1032. https://doi.org/10.1177/0899764013508009 arXiv:https://doi.org/10.1177/0899764013508009

[26] Raffaele Fabio Ciriello, Alexander Richter, and Gerhard Schwabe. 2018. Digital innovation. *Business & Information Systems Engineering* 60 (2018), 563–569.

[27] Roderic Crooks and Morgan Currie. 2021. Numbers will not save us: Agonistic data practices. *The Information Society* 37, 4 (2021), 201–213.

[28] Sajjad Dadkhah, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, and Ali A. Ghorbani. 2023. The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker. *IEEE Transactions on Computational Social Systems* (2023), 1–15. https://doi.org/10.1109/TCSS.2023.3322303

[29] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 525–534.

[30] danah boyd and Kate Crawford. 2012. Critical Questions For Big Data. *Information, Communication & Society* 15, 5 (2012), 662–679. https://doi.org/10.1080/1369118X.2012.678878 arXiv:https://doi.org/10.1080/1369118X.2012.678878

[31] Stevienna de Saille and Fabien Medvecky. 2016. Innovation for a steady state: a case for responsible stagnation. *Economy and Society* 45, 1 (2016), 1–23.

[32] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 473–484.

[33] Catherine DIgnazio and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press, Cambridge, Massachusetts and London, England.

[34] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. 2016. Social justice-oriented interaction design: Outlining key design strategies and commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. Association

[35] M. C. Elish and danah boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs* 85, 1 (2018), 57–80.

[36] Elizabeth Ellcessor. 2016. *Restricted access: Media, disability, and the politics of participation*. Vol. 6. NYU Press.

[37] Sheena Erete, Emily Ryou, Geoff Smith, Khristina Marie Fassett, and Sarah Duda. 2016. Storytelling with Data: Examining the Use of Data by Non-Profit Organizations. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1273–1283. https://doi.org/10.1145/2818048.2820068

[38] Better Evaluation. 2014. Rainbow Framework. https://www.betterevaluation.org/en/rainbow_framework

[39] Hayley I. Evans, Marisol Wong-Villacres, Daniel Castro, Eric Gilbert, Rosa I. Arriaga, Michaelanne Dye, and Amy Bruckman. 2018. *Facebook in Venezuela: Understanding Solidarity Economies in Low-Trust Environments*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173802

[40] Luis Fernandez-Luque and Muhammad Imran. 2018. Humanitarian health computing using artificial intelligence and social media: A narrative literature review. *International journal of medical informatics* 114 (2018), 136–142.

[41] Clayton Fink, Christine D Piatko, James Mayfield, Tim Finin, Justin Martineau, et al. 2009. Geolocating Blogs from Their Textual Content. In *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*. AAAI Press, 25–26.

[42] Kathleen Foody and Matt O Brien. 2023. Confusion as Musk's Twitter Yanks Blue checks from agencies. https://apnews.com/article/twitter-elon-musk-blue-checkmark-celebrities-544cfd66ed3a62f51a8a80c20e11ac5b

[43] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake news on Facebook and Twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[44] Michael C Gilbert. 2001. Disconnected: the first nonprofit email survey.

[45] Maya Goodwill, Mieke van der Bijl-Brouwer, and Roy Bendor. 2021. Beyond good intentions: Towards a power literacy framework for service designers. *International Journal of Design* 15, 3 (2021), 45–59.

[46] Saul Greenberg. 2006. Toolkits and Interface Creativity. *Multimedia Tools and Applications* 32, 2 (2006), 139–159. https://doi.org/10.1007/s11042-006-0062-y

[47] Darrene Hackler and Gregory D Saxton. 2007. The strategic use of information technology by nonprofit organizations: Increasing capacity and untapped potential. *Public administration review* 67, 3 (2007), 474–487.

[48] Eszter Hargittai. 2020. Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review* 38, 1 (2020), 10–24. https://doi.org/10.1177/0894439318788322 arXiv:https://doi.org/10.1177/0894439318788322

[49] Ellie Harmon, Chris Bopp, and Amy Voida. 2017. The design fictions of philanthropic IT: Stuck between an imperfect present and an impossible future. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 7015–7028.

[50] Ellie Harmon, Matthias Korn, and Amy Voida. 2017. Supporting everyday philanthropy: Care work in situ and at scale. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 1631–1645.

[51] Ron Haskins and Jon Baron. 2011. Building the connection between policy and evidence. http://coalition4evidence.org/wp-content/uploads/2011/09/Haskins-Baron-paper-on-fed-evid-based-initiatives-2011.pdf

[52] Starr Roxanne Hiltz, Jane A Kushma, and Linda Plotnick. 2014. Use of Social Media by US Public Sector Emergency Managers: Barriers and Wish Lists. *ISCRAM* 10, 2.1 (2014), 3122–4005.

[53] Nathan O Hodas, Greg Ver Steeg, Joshua Harrison, Satish Chikkagoudar, Eric Bell, and Courtney D Corley. 2015. Disentangling the lexicons of disaster response in twitter. In *Proceedings of the 24th International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA, 1201–1204.

[54] Christina Holt, Stephen Fawcett, Vincent Francisco, Jerry Schultz, Bill Berkowitz, and Tom Wolff. 1994. The Community Tool Box. https://ctb.ku.edu/en/toolkits

[55] J Brian Houston, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A McElderry, et al. 2015. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters* 39, 1 (2015), 1–22.

[56] Amanda L Hughes, Lise AA St. Denis, Leysia Palen, and Kenneth M Anderson. 2014. Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1505–1514.

[57] Kristina Jaskyte. 2012. Exploring potential for information technology innovation in nonprofit organizations. *Journal of Technology in Human Services* 30, 2 (2012), 118–127.

[58] Andrea Jimenez, Deborah Delgado, Roger Merino, and Alejandro Argumedo. 2022. A decolonial approach to innovation? Building paths towards Buen Vivir. *The Journal of Development Studies* 58, 9 (2022), 1633–1650.

[59] Antonios Karteris, Georgios Tzanos, Lazaros Papadopoulos, Konstantinos Demestichas, Dimitrios Soudris, Juliette Pauline Philibert, and Carlos López Gómez. 2022. A Methodology for Enhancing Emergency Situational Awareness through Social Media. In *Proceedings of the 17th International Conference on Availability, Reliability and Security* (Vienna, Austria) *(ARES '22)*. Association for Computing Machinery, New York, NY, USA, Article 130, 7 pages. https://doi.org/10.1145/3538969.3544418

[60] Naveena Karusala, Jennifer Wilson, Phebe Vayanos, and Eric Rice. 2019. Street-Level Realities of Data Practices in Homeless Services Provision. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 184 (nov 2019), 23 pages. https://doi.org/10.1145/3359286

[61] Michael Kaschesky, Pawel Sobkowicz, José Miguel Hernández Lobato, Guillaume Bouchard, Cedric Archambeau, Nicolas Scharioth, Robert Manchin, Adrian Gschwend, and Reinhard Riedl. 2013. Bringing Representativeness into Social Media Monitoring and Analysis. (2013), 2003–2012. https://doi.org/10.1109/HICSS.2013.120

[62] Theo Keane, Brenton Caffin, Michael Soto, Ayush Chauhan, Rikta Krishnaswamy, Geke van Dijk, and Megha Wadhawan. 2014. DIY Development Impact and You: Practical tools to trigger and support social innovation. https://www.nesta.org.uk/toolkit/diy-toolkit/

[63] Christopher M Kelty. 2017. The Participatory Development Toolkit.

[64] Vera Khovanskaya, Lynn Dombrowski, Ellie Harmon, Matthias Korn, Ann Light, Michael Stewart, and Amy Voida. 2018. Designing against the Status Quo. *Interactions* 25, 2 (feb 2018), 64–67. https://doi.org/10.1145/3178560

[65] Vera Khovanskaya, Phoebe Sengers, and Lynn Dombrowski. 2020. Bottom-Up Organizing with Tools from On High: Understanding the Data Practices of Labor Organizers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376185

[66] Ulrike Krause. 2013. Innovation: the new big push or the post-development alternative? *Development* 56, 2 (2013), 223–226.

[67] Mark Latonero and Irina Shklovski. 2011. Emergency management, Twitter, and social media evangelism. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 3, 4 (2011), 1–16.

[68] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S Hopkins, and Nina Kruschwitz. 2011. Big data, analytics and the path from insights to value. *MIT sloan management review* 52, 2 (2011), 21–32.

[69] Christopher A. Le Dantec, Mariam Asad, Aditi Misra, and Kari E. Watkins. 2015. Planning with Crowdsourced Data: Rhetoric and Representation in Transportation Planning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1717–1727. https://doi.org/10.1145/2675133.2675212

[70] Amy Voida Lehn M. Benjamin and Chris Bopp. 2018. Policy fields, data systems, and the performance of nonprofit human service organizations. *Human Service Organizations: Management, Leadership & Governance* 42, 2 (2018), 185–204. https://doi.org/10.1080/23303131.2017.1422072 arXiv:https://doi.org/10.1080/23303131.2017.1422072

[71] Yannan Li and Amy Voida. 2022. Nonprofit Organizations' Dialogic Use of Social Media: Principles and Practice. *Journal of Nonprofit & Public Sector Marketing* 36, 1 (2022), 1–26.

[72] Ann Light. 2011. HCI as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with computers* 23, 5 (2011), 430–438.

[73] Yanni Alexander Loukissas. 2019. *All Data Are Local: Thinking Critically in a Data-Driven Society.* The MIT Press, Cambridge, Massachusetts.

[74] Mirca Madianou. 2019. Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises. *Social Media + Society* 5, 3 (2019), 2056305119863146. https://doi.org/10.1177/2056305119863146 arXiv:https://doi.org/10.1177/2056305119863146

[75] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2021. Where Is This Tweet From? Inferring Home Locations of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media* 6, 1 (Aug. 2021), 511–514. https://doi.org/10.1609/icwsm.v6i1.14299

[76] Carleen Maitland, Jean-Laurent Martin, Maria Gabriela Urgiles Bravo, and Alex Bertram. 2022. A Qualitative Difference: Integrating Qualitative Data into Humanitarian Response Operations. In *Proceedings of the Twelfth International Conference on Information and Communication Technologies and Development* (Seattle, USA) *(ICTD '22)*. Association for Computing Machinery, New York, NY, USA, 25 pages.

[77] Jyoti Mann. 2022. Layoffs, long hours and RTO. https://www.businessinsider.com/elon-musk-twitter

[78] Nicolai Marquardt, Steven Houben, Michel Beaudouin-Lafon, and Andrew D. Wilson. 2017. HCITools: Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 624–627. https://doi.org/10.1145/3027063.3027073

[79] Nicolai Marquardt, Steven Houben, Michel Beaudouin-Lafon, and Andrew D. Wilson. 2017. HCITools: Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 624–627. https://doi.org/10.1145/3027063.3027073

[80] Aaron Martin, Gargi Sharma, Siddharth Peter de Souza, Linnet Taylor, Boudewijn van Eerd, Sean Martin McDonald, Massimo Marelli, Margie Cheesman, Stephan Scheel, and Huub Dijstelbloem. 2022. Digitisation and Sovereignty in Humanitarian Space: Technologies, Territories and Tensions. *Geopolitics* 0, 0 (2022), 1–36. https://doi.org/10.1080/14650045.2022.2047468 arXiv:https://doi.org/10.1080/14650045.2022.2047468

[81] Shannon Mattern. 2021. Unboxing the Toolkit.

[82] EE McPherson. 2015. *ICTs and Human Rights Practice: A Report Prepared for the UN Special Rapporteur on Extrajudicial, Summary, or Arbitrary Executions.* University of Cambridge Centre of Governance and Human Rights, Cambridge,England. https://doi.org/10.17863/CAM.16807

[83] Patrick Meier. 2015. *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response.* CRC Press, Inc., USA.

[84] Cecelia Merkel, Umer Farooq, Lu Xiao, Craig Ganoe, Mary Beth Rosson, and John M Carroll. 2007. Managing technology use and learning in nonprofit community organizations: Methodological challenges and opportunities. In *Proceedings of the 2007 symposium on Computer human interaction for the management of information technology.* Association for Computing Machinery, New York, NY, USA, 8–es.

[85] Cecelia Bridget Merkel, Mike Clitherow, Umer Farooq, Lu Xiao, Craig Harvey Ganoe, John M Carroll, and Mary Beth Rosson. 2005. Sustaining computer use and learning in community computing contexts: Making technology part of who they are and what they do. *The Journal of Community Informatics* 1, 2 (2005), 158–174.

[86] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 115 (Oct. 2020), 25 pages. https://doi.org/10.1145/3415186

[87] Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. 2022. Documenting Data Production Processes: A Participatory Approach for Data Work. *[forthcoming]* 1, 1 (2022), 34.

[88] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. https://api.semanticscholar.org/CorpusID:5959482

[89] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013), 3111–3119.

[90] Michael Muller, Cecilia Aragon, Shion Guha, Marina Kogan, Gina Neff, Cathrine Seidelin, Katie Shilton, and Anissa Tanweer. 2020. Interrogating Data Science. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 467–473. https://doi.org/10.1145/3406865.3418584

[91] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 94, 16 pages. https://doi.org/10.1145/3411764.3445402

[92] Steven Lee Myers, Stuart A. Thompson, and Tiffany Hsu. 2023. The consequences of Elon Musk's ownership of x. https://www.nytimes.com/interactive/2023/10/27/technology/twitter-x-elon-musk-anniversary.html

[93] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (2019), 13. https://doi.org/10.3389/fdata.2019.00013

[94] Lucy Pei, Benedict Salazar Olgado, and Roderic Crooks. 2022. Narrativity, Audience, Legitimacy: Data Practices of Community Organizers. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 328, 6 pages. https://doi.org/10.1145/3491101.3519673

[95] Adrian Petterson, Keith Cheng, and Priyank Chandra. 2023. Playing with Power Tools: Design Toolkits and the Framing of Equity. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 392, 24 pages. https://doi.org/10.1145/3544548.3581490

[96] Fabian Pformüller, Nico Luchsinger, and Sascha Mombartz. 2017. The community canvas. https://community-canvas.org/

[97] James Pierce, Sarah Fox, Nick Merrill, and Richmond Wong. 2018. Differential vulnerabilities and a diversity of tactics: What toolkits teach us about cybersecurity. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[98] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19, 1 (Dec. 2013). https://doi.org/10.5210/fm.v19i1.4848

[99] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[100] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 236, 13 pages. https://doi.org/10.1145/3411764.3445604

[101] Lester Salomon, Wojciech Sokolowski, and Regina List. 2003. Global civil society: an overview.

[102] Anita Saroj and Sukomal Pal. 2020. Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction* 48 (2020), 101584.

[103] Laura Scheepmaker, Kay Kender, Christopher Frauenberger, and Geraldine Fitzpatrick. 2021. Leaving the Field: Designing a Socio-Material Toolkit for Teachers to Continue to Design Technology with Children. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 145, 14 pages. https://doi.org/10.1145/3411764.3445462

[104] Stefan Schmager and Sonia Sousa. 2021. A Toolkit to Enable the Design of Trustworthy AI. In *HCI International 2021 - Late Breaking Papers: Multimodality, eXtended Reality, and Artificial Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Jessie Y. C. Chen, Gino Fragomeni, Norbert Streitz, Shin'ichi Konomi, Helmut Degen, and Stavroula Ntoa (Eds.). Springer International Publishing, Cham, 536–555.

[105] Rajeev Sharma, Sunil Mithas, and Atreyi Kankanhalli. 2014. Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems* 23, 4 (2014), 433–441.

[106] Hong Shen, Wesley H. Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 850–861. https://doi.org/10.1145/3442188.3445971

[107] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 440–451. https://doi.org/10.1145/3531146.3533110

[108] Chris Stokel-Walker. 2023. Twitter's $42,000-per-month API prices out nearly everyone. https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/

[109] Katie G. Tanaka and Amy Voida. 2016. Legitimacy Work: Invisible Work in Philanthropic Crowdfunding. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4550–4561. https://doi.org/10.1145/2858036.2858110

[110] Charlotte Tang, Yunan Chen, Bryan C. Semaan, and Jahmeilah A. Roberson. 2015. Restructuring Human Infrastructure: The Impact of EHR Deployment in a Volunteer-Dependent Clinic. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 649–661. https://doi.org/10.1145/2675133.2675277

[111] Martina Tazzioli. 2022. Extract, Datafy and Disrupt: Refugees' Subjectivities between Data Abundance and Data Disregard. *Geopolitics* 27, 1 (2022), 70–88. https://doi.org/10.1080/14650045.2020.1822332 arXiv:https://doi.org/10.1080/14650045.2020.1822332

[112] Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 505–514. https://doi.org/10.1609/icwsm.v8i1.14517

[113] Tweeter, Inc. 2022. What's new with v2. https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api. Accessed: 2022-04-11.

[114] Sarah Vieweg, Carlos Castillo, and Muhammad Imran. 2014. *Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters*. Springer International Publishing, Cham, 444–461. https://doi.org/10.1007/978-3-319-13734-6_32

[115] Amy Voida. 2011. Shapeshifters in the voluntary sector: exploring the human-centered-computing challenges of nonprofit organizations. *interactions* 18, 6 (2011), 27–31.

[116] Amy Voida, Lynn Dombrowski, Gillian R. Hayes, and Melissa Mazmanian. 2014. Shared values/conflicting logics: working around e-government systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3583–3592. https://doi.org/10.1145/2556288.2556971

[117] Amy Voida, Ellie Harmon, and Ban Al-Ani. 2011. Homebrew Databases: Complexities of Everyday Information Management in Nonprofit Organizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 915–924. https://doi.org/10.1145/1978942.1979078

[118] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[119] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.

[120] Marisol Wong-Villacres, Cristina M. Velasquez, and Neha Kumar. 2017. Social Media for Earthquake Response: Unpacking Its Limitations with Care. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 112 (2017), 22 pages. https://doi.org/10.1145/3134747

[121] Dingqi Yang, Daqing Zhang, Korbinian Frank, Patrick Robertson, Edel Jennings, Mark Roddy, and Michael Lichtenstern. 2014. Providing real-time assistance in disaster relief by leveraging crowdsourcing power. *Personal and Ubiquitous Computing* 18 (2014), 2025–2034.

[122] Max Zahn. 2022. A timeline of Elon Musk's tumultuous Twitter acquisition. https://abcnews.go.com/Business/timeline-elon-musks-tumultuous-twitter-acquisition-attempt/story?id=86611191

[123] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How Do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 022 (May 2020), 23 pages. https://doi.org/10.1145/3392826

[124] Ziyuan Zhong. 2020. A tutorial on Fairness in Machine Learning. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

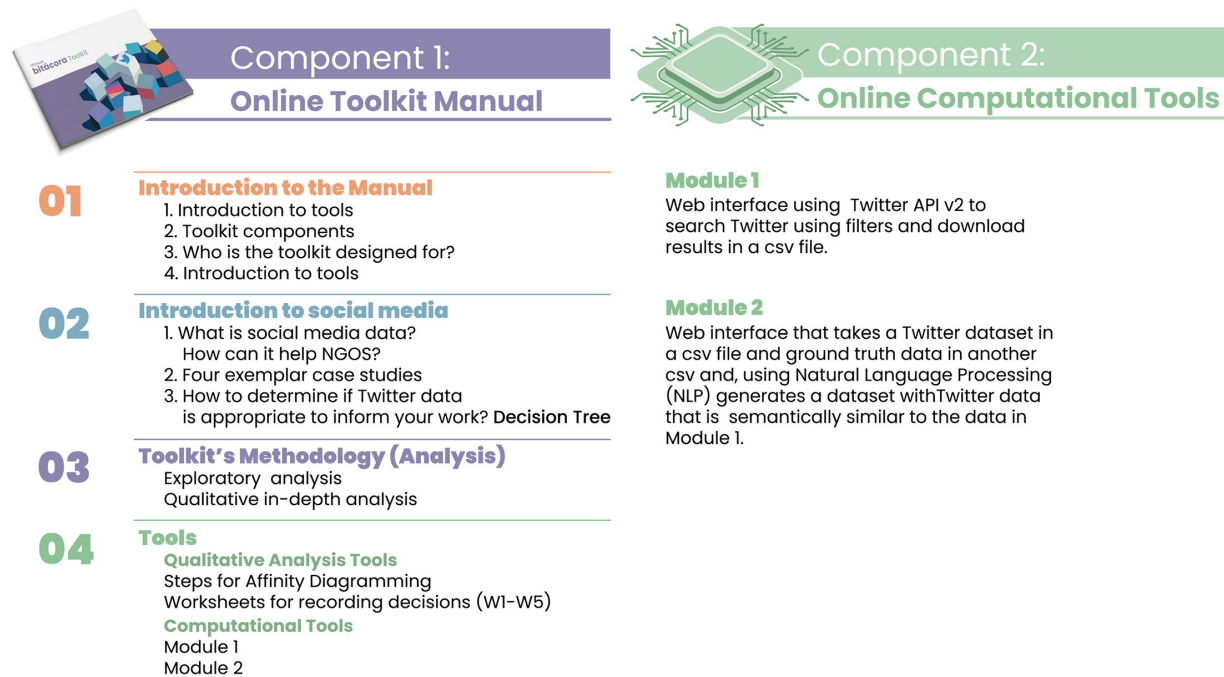## Appendix A   Details on the Bitacora's Components

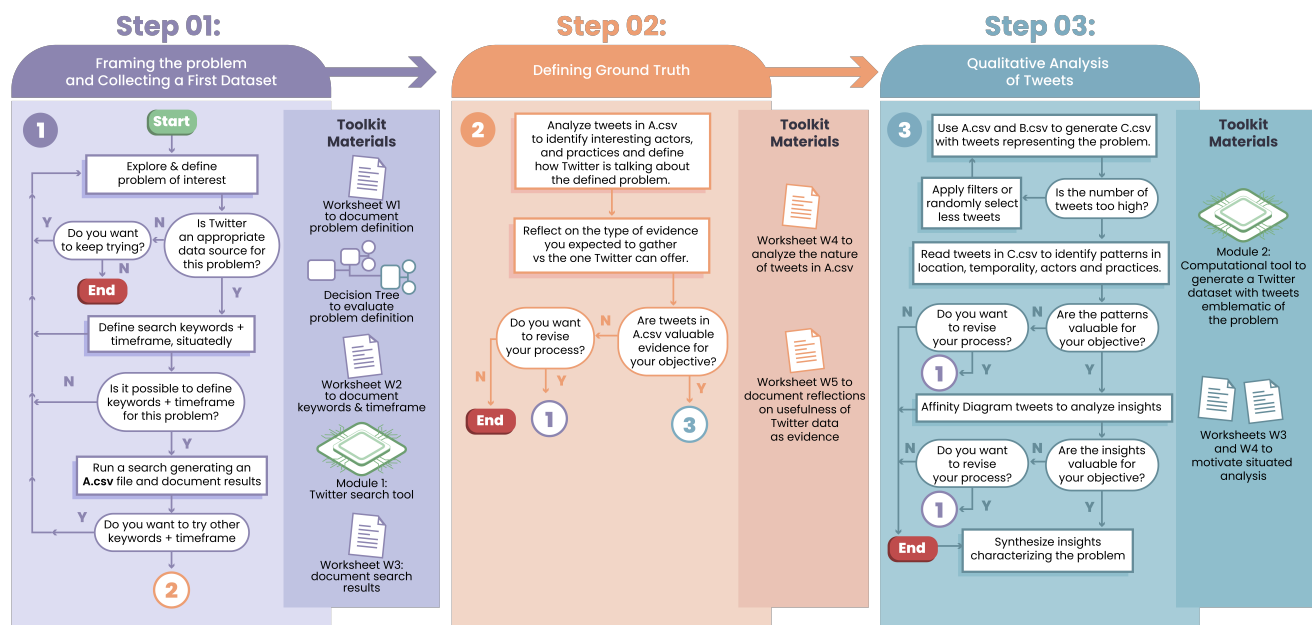**Figure 1: Overview of the toolkit's components**

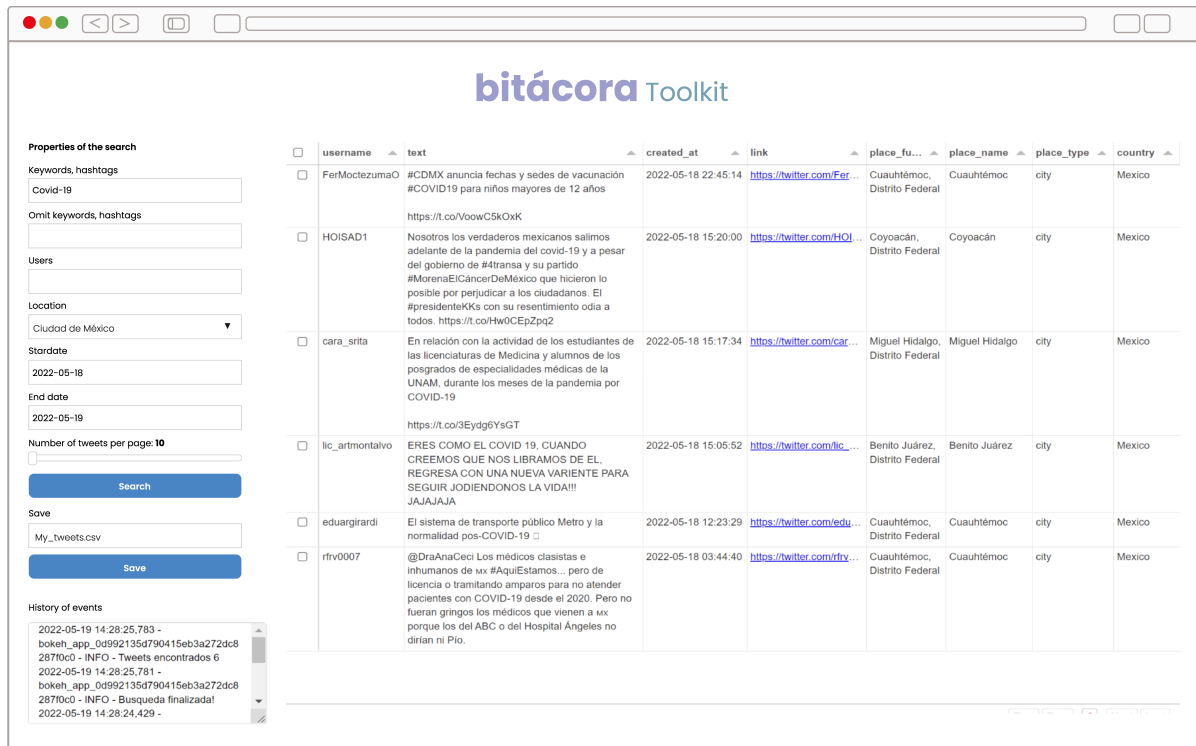**Figure 2: Diagram of the toolkit's methodology**

Figure 3: User Interface of Module 1
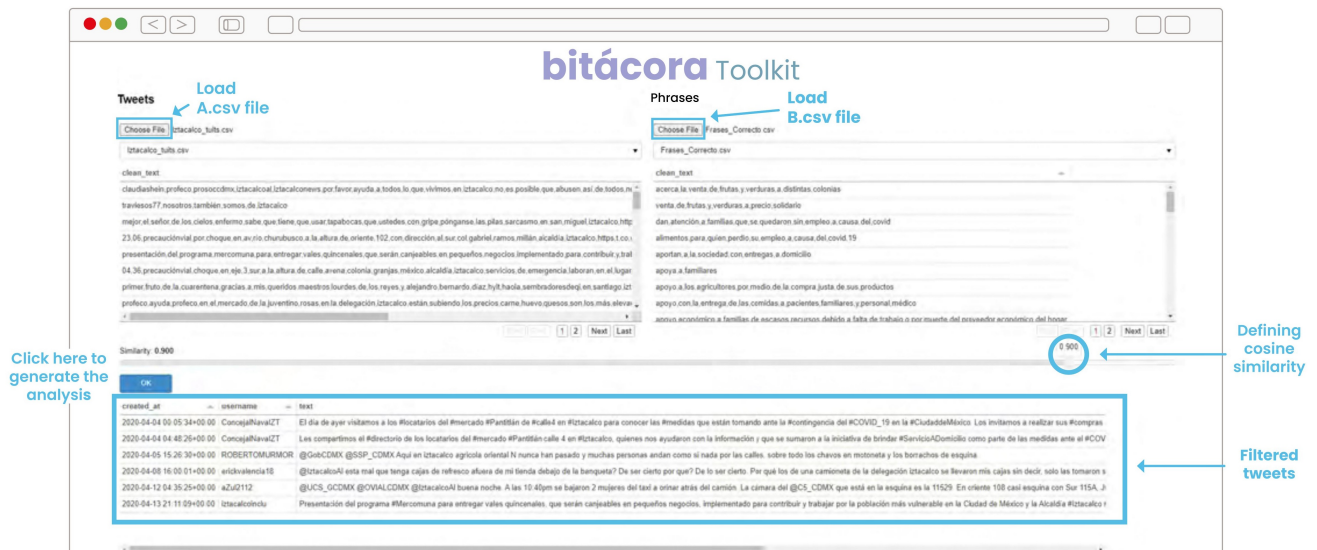
Figure 4: User Interface of Module 2

Appendix B   Details about the Participants

ID	Org	Role	Background	Problem Exploration Methods' Expertise
1	Accelerator Lab	Researcher	B.A in Architecture and MSc in Sustainable Development	Experience in community building through participatory methods and conducting ethnographic studies. Expertise in qualitative methods and limited experience with quantitative methods.
2	Accelerator Lab	Intern	PhD candidate in sociology	Expertise conducting in-depth interviews and longitudinal studies. Limited experience with quantitative methods.
3	Accelerator Lab	Communications expert	B.A in Journalism	Expertise in content creation and social media content manager. Expertise in qualitative methods. Limited experience with quantitative methods.
4	Accelerator Lab	Researcher	B.A in Industrial Design and MA in Future Studies	Experience with ethnographic methods. Limited experience with quantitative methods.
5	Accelerator Lab	Intern	B.S. in Economics and MSc in Public Policy	Experience using official sources of information such as surveys and official indicators. Expertise in quantitative methods. Limited experience with qualitative methods.
6	CIEP	Researcher	PhD in Public Policy	Her research focuses on public spending and the care economy. Experience with quantitative methods. Limited experience with qualitative methods.
7	CIEP	Researcher	B.S. in Economics	Her research focuses on infrastructure and public investment. Experience with quantitative methods, and limited experience with qualitative methods.
8	CIEP	Researcher	B.S. in Economics	His research focuses in income and public debt. Experience with quantitative methods. Limited experience with qualitative methods.
9	SocialTic	Community building coordinator	B.S. in Economics	Extensive experience analyzing social media data from different platforms using mixed-methods. Experience with quantitative methods. Limited experience with qualitative methods.
10	SocialTic	Community building coordinator	B.A in Communication and Media	Extensive experience analyzing social media data from different platforms using mixed-methods. Experience with qualitative methods. Limited experience with quantitative methods.

Table 1: Participants' description: 1) IDs, 2) Organization name, 3) Role within the organization, 4) Formal training, and 5) Methods of expertise.

B.1   Details on the Projects' Framing

ID	Topics	Initial Framing of the Problem	Final Framing of the Problem
1	Resilience, risks and gender	How do the government and civil organizations discuss these topics?	Examine if there were discussions around gender disparities in response to the earthquake in Mexico in 2017. Who talks about this? Did any government agency develop target actions to address gender inequality?
2	Community response to natural disasters	Catalog of practices and responses of organizations and government for attention or recovery to communities after a natural disaster. Choose five iconic cases where there has been a catastrophe in the country.	Examined and compared the citizen organization and communication in two crises due to natural disasters. The participant examined the response to the earthquake in Mexico City in 2017 and the landslide in 2021 in the Cerro del Chiquihuite located north of Mexico City.
3	Gentrification, increased rents in Mexico, and displacement of local people by foreigners	Examine if the increase in the cost of rent is happening systemically and if it has the potential of becoming a social problem.	Narrowed the scope of the search to examine the trends of rental prices in Mexico City.
4	Mining companies in Mexico	Recently, the Supreme Court of Mexico revoked concessions to Canadian mining companies. What is being said about it?	Narrowed the search following the feedback received and focused on searching information about a community in the state of Puebla where the mining concession of a Canadian company was revoked.
5	Consultation with indigenous communities	Examine if actors from the private sector or government had discussed this topic and their respective position.	Reduce the scope of the search to only examine the opinion of private sector actors regarding indigenous communities.
6, 7, 8	The economic budget and package in Mexico	Answer the following questions: Who are the stakeholders involved in the conversations around the economic package? Are they consuming the reports produced by the organization? If yes, how are they using the organization's reports?	Participants did not changed their framing.
9, 10	Feminicides in Mexico	Examine the discourse of feminicides in Mexico. Evaluate whether Twitter can be an alternative source of information to document cases of feminicide. Using Twitter data, to examine how media coverage of femicides has evolved over time.	Participants did not changed their framing.

Table 2: Description of the initial and final framing of topics explored by participants.

Appendix C   Decision Tree

To simplify practitioners' decision process on whether or not to use social media data to inform their work, the manual includes a decision tree that makes it easy to rule out issues that cannot be examined with social media data.

How to decide if Twitter data is appropriate to inform your work?

Use the following decision tree to decide whether Twitter data is the most appropriate source of information to inform your work.
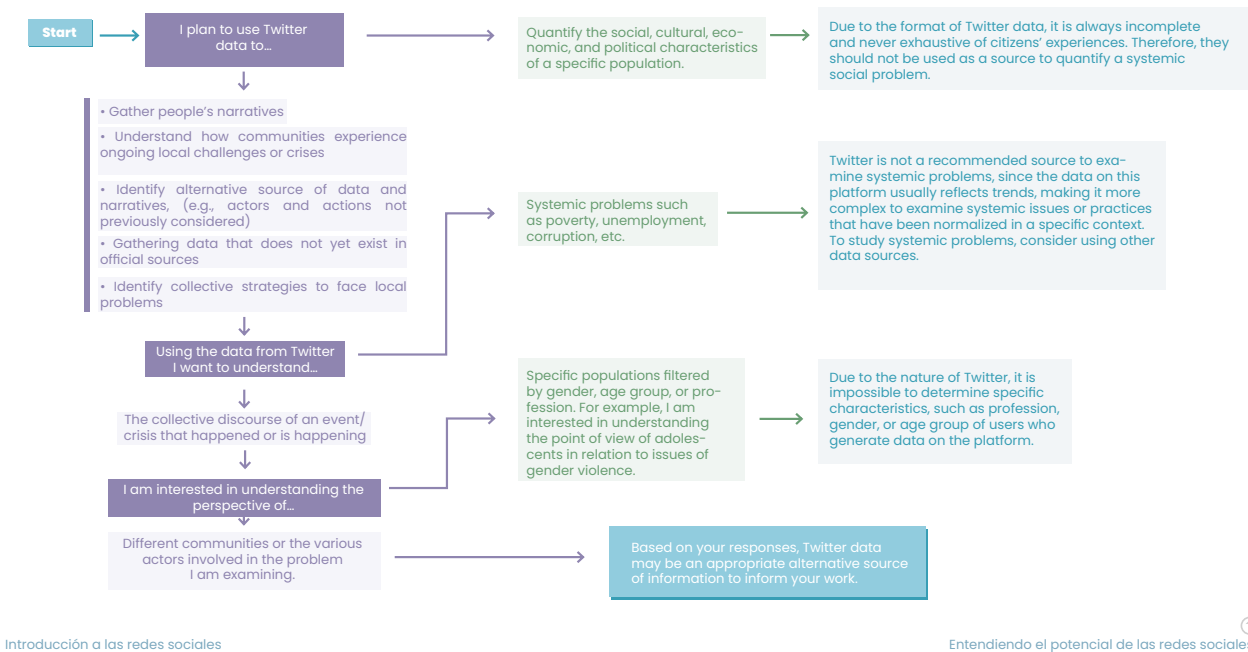

Start

I plan to use Twitter data to...

- Gather people's narratives
- Understand how communities experience ongoing local challenges or crises
- Identify alternative source of data and narratives, (e.g., actors and actions not previously considered)
- Gathering data that does not yet exist in official sources
- Identify collective strategies to face local problems

Using the data from Twitter I want to understand...

The collective discourse of an event/ crisis that happened or is happening

I am interested in understanding the perspective of...

Different communities or the various actors involved in the problem I am examining.

Quantify the social, cultural, economic, and political characteristics of a specific population.

Due to the format of Twitter data, it is always incomplete and never exhaustive of citizens' experiences. Therefore, they should not be used as a source to quantify a systemic social problem.

Systemic problems such as poverty, unemployment, corruption, etc.

Twitter is not a recommended source to examine systemic problems, since the data on this platform usually reflects trends, making it more complex to examine systemic issues or practices that have been normalized in a specific context. To study systemic problems, consider using other data sources.

Specific populations filtered by gender, age group, or profession. For example, I am interested in understanding the point of view of adolescents in relation to issues of gender violence.

Due to the nature of Twitter, it is impossible to determine specific characteristics, such as profession, gender, or age group of users who generate data on the platform.

Based on your responses, Twitter data may be an appropriate alternative source of information to inform your work.

Introducción a las redes sociales

Entendiendo el potencial de las redes sociales

Figure 5: Decision tree from the manual

Appendix D Description of the Worksheets

The worksheets encourage practitioners to reflect on the definitions of evidence in their work and help users document and reflect on their decisions when associating social media data with a specific issue.

Worksheet 1 - Problem Definition Template: This worksheet guides practitioners in defining the problem to be examined using data from Twitter, and it consists of five sections.

(1) **Description and expectations:** This section is at the center of the template, and it asks for a short description of the problem and users' expectations using data from Twitter.

(2) **Context (place and time):** This section asks three questions: 1) where is this problem happening?, 2) is it a widespread and generalized problem or a problem of a specific area?, and 3) what is the timeframe of the problem? The intention of including these questions is to encourage the user to reflect on what information they have about the problem to be examined and the problem characteristics in terms of place and time.

(3) **Communities:** This section asks for the communities involved in the problem to be examined.

(4) **Evidence and data:** This section asks the following questions: 1) what evidence is currently available on the problem? 2) what kind of evidence do you expect to find? 3) what type of data is currently available?

(5) **Objectives:** What is the motivation for using Twitter data? List the questions that you seek to answer with the exploration of Twitter

Figure 6: Worksheet 1 - Problem Definition Template

Worksheet 2 - Keyword Documentation: This worksheet aimed to help participants document the keywords that characterize the problem practitioners wanted to explore on Twitter. The template has five columns of possible keyword categories users decide to register. The categories are *(i)* description of the problem, *(ii)* names of organizations involved, *(iii)* affected communities, *(iv)* Twitter accounts of actors involved, and *(vi)* Hashtags.

Keyword Documentation

Use this template to record the keywords that characterize the issue you want to explore on Twitter.

Manual
bitácora
Toolkit

Name of the project: _____   Date: _____

Description of the problem	Names of organizations involved	Affected communities	Twitter accounts of actors involved	Hashtags
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Description:

[ ]

Figure 7: Worksheet 2 - Keyword Documentation Template

Worksheet 3 - Exploratory Search Documentation Template This template is to help users document their observations from exploratory searches. In the manual, I suggested practitioners use the worksheet to keep track of which combinations of words yield the most unusual or appropriate content that contributes to the understanding of the examined problem. The template is divided into seven columns with the categories of aspects that I recommend to document during the first observations. These categories are: 1) place, 2) time, 3) the combination of keywords, 4) number of results, 5) observations, 6) actors/Twitter accounts, and 7) examples of tweets.

Exploratory Search Documentation Template

Use this template to document your observations from exploratory searches.

Manual
bitácora
Toolkit

Name of the project: _____

Date: _____

- **Documentation of exploratory searches** -

ID	Place	Time	Combination of keywords	Number of results	Observations	Actors/ Twitter accounts	Examples of tweets
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							

Description:

Figure 8: Worksheet 3 - Exploratory Search Documentation Template

Worksheet 4 - Findings and Observations Documentation Template The purpose of this template is to document and synthesize the most relevant conclusions of the initial examination. Ideally, this template needs to be completed after analyzing the findings of the initial explorations on Twitter in the worksheet 3. On the left side of the template, there are three small boxes where practitioners can enter the project's name, a brief description, and a summary of the focus of the exploration, for example, listing some of the search words. The purpose of these three boxes is to summarize the project. In the template, there is also a text box divided into three sections:

(1) **List relevant actors:** This section asks to list any new actors or communities involved in the problem examined that might have been found on the exploration of Twitter. The form asks 1) who are they? 2) what is interesting about them? 3) what is their role?
(2) **List initial findings:** This field asks to describe how the topic of interest is discussed on Twitter and recommends recording examples of tweets that are illustrative of the findings, as this facilitates the communication of results.
(3) **List relevant actions or practices:** This section asks practitioners to document any new identified perspectives or interesting initiatives that contribute to the problem examined and to consider answering the following questions:
  (a) What is the purpose of these initiatives?
  (b) What distinguishes them from existing initiatives?

Findings and Observations Documentation Template

Use this template to document the main findings of the initial exploration.

Manual
bitácora
Toolkit

Project's name

Brief description

Summary of the goal
of the exploration

List interesting observations
List relevant actors

If you identified new actors or
communities involved in the
problem examined, list them here:
Who are they?
What is interesting about them?
What is your role?

List initial findings
If Briefly describe how the topic
of interest is discussed on Twitter

List interesting observations
List relevant actions or practices

If you identified new perspectives or
interesting initiatives that contribute
to the problem examined,
describe and list them here:
What is the purpose of these initiatives?
What distinguishes them
from existing initiatives?

Figure 9: Worksheet 4 - Findings and Observations Documentation Template

Worksheet 5 - Next Steps Documentation Template The purpose of this template is to encourage practitioners to reflect on the observations collected in the exploratory searches, assess whether the narrative observed on Twitter is of interest for further analysis, and determine the next steps. After participants complete the template, they would need to decide whether or not it is advisable to continue using data from Twitter to report the problem examined. Similar to worksheet 4 on the left side, there are three small text boxes where participants can synthesize the project focus and the keywords they used during the exploration.

Next Steps Documentation Template

Use this template to document your observations from exploratory searches.

Manual
bitácora
Toolkit

Project's name

Brief description

Summary of the goal
of the exploration

Relationship between Twitter findings and prior knowledge

How are the observations found on Twitter related or different from what you knew about the problem before?
What's new?
What information or patterns are repeated?

Potential limitations

What limitations do you anticipate? For example, not finding information about a specific place or population.
What are some questions or concerns?

Do you consider it necessary to carry out a more in-depth search?

Yes

How do you think Twitter data contribute to understanding the research problem you are examining?

No

Why?

Figure 10: Worksheet 5 - Next Steps Documentation Template