2022 IEEE 30th International Requirements Engineering Conference Workshops (REW) | 978-1-6654-6000-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/REW56159.2022.00031

Phish Finders: Crowd-powered RE for anti-phishing training tools

Holly Rosser, Maylene Mayor, Adam Stemmler, Vinod Ahuja, Andrea Grover and Matthew Hale
University of Nebraska Omaha
Omaha, USA
Email: hrosser, mmayor, astemmler, vahuja, andreagrover, mlhale@unomaha.edu

***Abstract*—Many organizations use internal phishing campaigns to gauge awareness and coordinate training efforts based on those findings. Ongoing content design is important for phishing training tools due to the influence recency has on phishing susceptibility. Traditional approaches for content development require significant investment and can be prohibitively costly, especially during the requirements engineering phase of software development and for applications that are constantly evolving. While prior research primarily depends upon already known phishing cues curated by experts, our project, Phish Finders, uses crowdsourcing to explore phishing cues through the unique perspectives and thought processes of everyday users in a realistic yet safe online environment, Zooniverse. This paper contributes qualitative analysis of crowdsourced comments that identifies novel cues, such as formatting and typography, which were identified by the crowd as potential phishing indicators. The paper also shows that crowdsourcing may have the potential to scale as a requirements engineering approach to meet the needs of content labeling for improved training tool development.**

***Index Terms*—citizen science, Zooniverse, crowdsourcing, cybersecurity, phishing**

I. Introduction

Internal phishing campaigns are commonly used by organizations to gauge awareness and coordinate training efforts [1]. These campaigns deliver targeted phishing emails within organizational email tools and then, based on the findings, route vulnerable users towards corrective training. Despite these efforts, large numbers of targets continue to be successfully exploited by phishing tactics [2]. As folks continue to be phished, researchers have sought to better understand why people click links and how sophisticated phishing attacks misdirect users, particularly those with the most internet literacy [3]. Findings from that research have led to a wave of new phishing training tools [4], [5]. These tools rely on labeled phishing content that users can view, interact with, and learn from in educational settings.

Content design is important for educational phishing training tools because these tools need to provide recent, varied phishing samples to trainees over time as they build recognition skills. Content recency, i.e., how fresh content is to the current time, trends, and news, has been shown to have a strong influence on phishing susceptibility and cognitive cues of trust and suspicion [6]. In short, legitimate content from 2018 can easily appear "phishy" to users in 2022. For developers building training tools, mechanisms for generating "fresh" content and then labeling it to identify phishing cues are core functional requirements. Given the semantic difficulty of labeling complex cues in online content, content generation for phishing tool support has yet to be automated [7].

During the requirements engineering phase of training tool development, software engineers need to identify sustainable approaches for sourcing, labeling, and importing recent phishing content into their tools. After a content management approach is identified, it needs to be deployed on an ongo-ing basis throughout the life of the application. Currently, addressing content generation recency and labeling requirements means gathering real phishing samples from online repositories and then engaging cybersecurity experts to label the data by hand before it is imported into a tool's content database. Expertly labeled content is very expensive and is not sustainable over the life of the software. The expense is most felt early in the development lifecycle, particularly during the requirements engineering phase when patterns of labeling content are not well established, and then continues after the deployment phase, as content must be constantly kept up-to-date and expert labelers come and go. While machine learning could partially support this process for well-established cues, parallel algorithms would be required for detecting novel cues as phishing techniques evolve. In addition, creating an ML training corpus would require the same foundational, ongoing effort as the labeled content for training tools, and may require an even larger training corpus to ensure adequate performance, making it an inefficient option for generating labeled content in this context. We believe that crowdsourcing can significantly reduce this cost, while improving the quality and recency of the content used in training tools.

II. Related Work

Crowdsourcing is increasingly used as an information source for software requirements that are constantly evolving and that do not have prototypical examples of stakeholders (i.e., not all trainees have comparable experience and knowledge, so we need to sample broad populations of trainees to better understand the range of perspectives to craft content for them) [8]–[10]. In this regard, wisdom of the crowd – the aggregate is more accurate than the sum of the individual solutions – is the overarching theory that supports the use of crowdsourcing for research [11]. Analysis of participant comments in this image classification project, Phish Finders, shows that collectively the crowd can do a good job (similar

2770-6834/22/$31.00 ©2022 IEEE
DOI 10.1109/REW56159.2022.00031

to an expert) at identifying known cues presented to them [12], and identify novel cue types that were not present in the taxonomy.

As the name suggests, the primary task in Phish Finders is to find the phish: participants in a crowdsourcing environment were tasked with identifying potentially suspicious features, called cues, in a collection of images that included both trustworthy and malicious content. Phishing cues have been studied and categorized; Staggs et al. [6] constructed a taxonomy that classifies and categorizes different phishing cues into different clusters according to the the technique phishers use. Their taxonomy was created by surveying the literature and findings from prior studies. The top-down literature-driven approach of creating the taxonomy provides a starting point for content development, but does not take the extreme variance, unique perspectives, and thought processes of everyday users into account and thus misses aspects of phishing cue detection researchers would never have given attention to. To broaden the perspectives for more effective requirements gathering, bottom-up emergent approaches can complement prior strategies. While some studies involving training people with anti-phishing tools, such as [13], have gathered trainee user feedback to further refine the taxonomy, the size of the user bases they draw from has remained small (i.e., a few hundred users).

Our project takes a bottom-up approach that seeks to broaden the base of participation to better find phishing cues by consulting the crowd to identify known and novel phishing cues and vectors. To focus our work, we proposed the following research question: *How can we use crowd-generated data to better satisfy content labeling requirements for anti-phishing training tools?*

III. Methods

To address the research question, we created an observational crowdsourcing study on a citizen science platform called Zooniverse where research projects, like Phish Finders, can engage volunteers in online image analysis. By attaining consensus across several volunteers for each image, projects on Zooniverse are capable of collecting robust data tailored to the needs of the project [14]. For Phish Finders, images were presented to volunteers for identification of phishing cues during January of 2021. The images, representing screenshots of various websites and emails, included both trustworthy and malicious content and were presented to participants for them to look over and assess. Two sets of images were used: a gold standard of 30 expert evaluated images containing 16 malicious and 14 trustworthy images, and another corpus of 1892 images containing 817 malicious and 1075 trustworthy images generated from websites spanning various sectors commonly used in phishing attacks. These sectors included banking, government, law enforcement, social networking, eCommerce, news, entertainment, and telecommunication sites. The websites were retrieved from the Internet Archive and extracted as an image using a screenshot program. Finally, a browser header was added for realism and to allow participants to identify suspicious domains in the browser address bar.

Participants were asked to first identify whether the content was trustworthy or malicious on a 5-point Likert scale. If a participant marked content as untrustworthy, they were provided an interface to mark up the image with bounding boxes to identify areas in the content that informed their decision (i.e., areas they thought were indicative of phishing.) For each marked area, participants were asked to apply a label for the most salient type of cue they saw; they could also apply multiple bounding boxes to the same region. Several pre-specified categories, which included Invalid Domain or Sender, Poor Spelling or Grammar, and Appeals to Action related to Greed, Urgency, or Authority, could be annotated by simply clicking on a menu in the markup interface. If the volunteer found other phishing cues that did not fit within these types, a bounding box labeled 'Other Phishy Findings' could be used instead. Use of this cue required the volunteer to enter a short textual description to best label the area of their selection.

This study focuses on the analysis of the 5,735 unstructured comments entered by participants as 'Other Phishy Findings'. To analyze the data, three researchers coded it using deductive content analysis techniques [15]. The codebook was derived from a taxonomy using several papers starting with Staggs et al. [6]. This work was further enhanced in Hale et al. [16] as sophistications (what makes content harder to identify) and degradations (what makes it easier to identify); both of which were further defined by Wethor [17]. This codebook included the pre-specified categories identified in Zooniverse's markup interface as well as additional sub-categories within the code groups of General Cues, Specific Cues, and User Context. In total, the researchers had 27 codes from which to work with. Coding was done in an iterative process with each of the three researchers working independently. When content was difficult to categorize, researchers would compare notes and come to consensus on appropriate codes. Inter-rater reliability using Krippendorff's Alpha calculation within the Atlas.ti qualitative analysis software package was initially measured at 78 percent. An additional meeting was held among the researchers to compare notes and argue to consensus on codes that remained fuzzy. Final inter-rater reliability of the data set was 82 percent.

Upon review of the content analysis results, there were three codes that had been disproportionately applied as the concepts they represent are inherently broad and therefore not very useful for our analysis. *Previous Interactions, Interpersonal Trust-Organizational (Interpersonal Trust)*, and *Engagement-Cognitive-Aesthetics (Aesthetics)* had each been applied to over 1,000 data points, while most other codes had 300 or fewer data points. To better understand the range of specific cues volunteers were reacting to, we then conducted inductive content analysis using three researchers to identify categories within these broadly-applied themes. Each of the three themes were independently and inductively coded by one researcher to surface relevant commonalities among the cues that volunteers identified. The final analysis of relevant themes within these codes was performed independently by one researcher.

IV. Results

The collection and labeling of phishing content, a key requirement for developing robust training tools, is currently incredibly manual and labor intensive. The process of curating images for annotation for the project on Zooniverse was more labor intensive than anticipated, requiring hundreds of hours over a period of four months. However, once Phish Finders was launched on the platform, 1,393 registered volunteers provided 28,425 classifications (individual annotations) on the data set of 1922 images in about five days, with an average of 15 participants annotating each image. The swift task completion alone confirmed our expectation that if the annotations were adequately robust, the labeling process would be more scalable with crowd participation, saving expert time and effort for more specialized functions.

The results we report here focus on new cue types identified through analysis of the comments, which further suggest that we can achieve a more comprehensive labeling of phishing cues, therefore better satisfying content requirements, when a diverse range of individuals contribute to the content labeling. The rest of this section discusses the new cue types identified in the three themes, Previous Interactions, Interpersonal Trust, and Aesthetics.

A. Categories in Previous Interactions

Previous Interactions were initially coded for 'Other Phishy Findings' comments where perceived irregularities identified by the participants were based on that participants' prior knowledge and experiences, i.e., the content defied expectations set by their previous interactions with a specific organization or type of organization. For example, "the IRS wouldn't ask for this kind of information" or "people in NYC don't give 1st floor as an address." In each of these cases, we first identified whether the participants' comment appeared to be context-specific (e.g., IRS) and then looked at verb usage to determine the level of prior knowledge and experience that went into their answer. The use of verbs like "wouldn't" and "don't" indicated that the participant was speaking from personal experiences.

In total, there were 1,098 comments identified in this theme with the majority of those comments indicating evidence of multiple codes. For example, the comment "First name personalization could/would be expected" would be coded with both *email cues-greeting* because of the context and *previous interactions* because of the participant's prior knowledge of e-mail protocols can be inferred. However, 301 comments were identified with only the theme Previous Interactions, so for these comments, we applied deeper inductive anal-ysis, which highlighted seven categories within the theme (Table I): *general web design, cybersecurity literacy, software experience, platform experience, organization expectations, topic experience,* and *uncustomary cues and content*. Table I provides the counts for each theme with examples.

Of particular interest given our research question were the comments coded as *general web design* and *organization expectations* where participants spoke to the recency of the content presented to them. *General web design* was defined as cues, words, or other page content was considered by the participant to be odd or uncommon as presented. Examples of this code included comments like, "not usual website content" and "odd to have on a homepage". *Organization expectations* were specifically mentioning organizations and what was perceived as customary for that particular company or type of organization. Comments like, "well-established site shouldn't be 'powered by random service'," or "a business with a domain using gmail for email? Uncommon!" are representative of this category.

TABLE I
Categories in Previous Interactions

Previous Interactions	Count	Example
Uncustomary content or cues	81	"Accounts never use card number as username"
General web design	80	"not usual website content"
Organization expectations	51	"PayPal would never ask for a photo of your ID"
Platform experience	24	"Google doesn't make you select email provider"
Software experience	23	"Flash is dead"
Topic experience	18	"Company messaging doesn't typically list an author"
Cybersecurity literacy	9	"Never give out passwords"

TABLE II
Categories in Interpersonal Trust

Interpersonal Trust	Count	Example
Failure to meet expectations	198	"No idea why, just tweaked my spidey senses... which could be completely wrong :)"
Requests for personal information	84	"Why is this info necessary for a refund request?"
Suspicious/ Untrustworthy	43	"I have no idea why anyone would want to do this or would trust this. It just seems very suspicious."
Google forms/docs/sheets	25	"Why would at&t use google docs for this? Dodgy"

B. Categories in Interpersonal Trust

The *Interpersonal Trust* code was applied to those comments that reflected the participants' perceived level of trust of an organization and their willingness to accept risk based on that level of trust. Examples of comments in this theme are "Word press site and an upload which may be OK but maybe not," and "I expect a reputable organisation to know my name or user name." This theme contained 2,142 comments and 889 of them had *interpersonal trust* as their only code.

To analyze this theme, we identified certain keywords that would be representative of how trust (or the lack thereof) could be inferred in participant comments. Words such as "trust" and "suspicious" were an obvious beginning, but looking deeper into the theme, it became apparent that the comments assigned here were full of question marks and interrogatories like "what" and "why" that provided clues about trustworthiness. A handful of categories within the theme Interpersonal Trust became apparent (Table II): *Failure*

TABLE III
CATEGORIES IN AESTHETICS

Aesthetics	Count	Example
Formatting	68	"Bad Formatting"
Google forms	45	"AT&T would not use Google Forms to capture information"
Font	38	"Does not look like a font that a reputable news source would choose"
Generic looking	33	"Very generic looking. Doesn't have AT&T's logo. AT&T doesn't use Google forms"
Odd	25	"This is odd"
Free	24	"Unlikely a university would be hosted on a free site builder"
Error	24	"The messages are irrelevant to the error"

to Meet Expectations, Generic, Random, Request for Personal Information, Suspicious/Untrustworthy, and *Google Forms*. However, as analysis progressed, the categories of Random and Generic, while in themselves representative of unique concerns for content developers, provided more credence to the overarching theme Failure to Meet Expectations, since users tend to have preconceived or learned expectations of online content, both in more general terms, along the lines of heuristics, as well as specific to content that is published by a known organization [1], [18]. The remaining categories, Google Forms and Request for Personal Information, overlapped conceptually with categories found in the other themes, Previous Interactions and Aesthetics, further discussed in the next section.

C. Categories in Aesthetics

The code *Aesthetics* is defined as the specific features of the interface, such as the screen layout and graphics, and the respondents' overall aesthetic impressions of the online content's attractiveness and sensory appeal. Comments in this theme focused on the appearance and general appeal of the interface, such as "the website layout is just off" or "stuff overlaps; bad design but claims to be tied to Facebook." Table III provides counts and examples for the most prevalent Aesthetics cues: in total, 2,068 participants comments were coded under this theme, with more than half, 1,089, having just the single code assigned to it.

As with the other results, there were similarities in the phishing cues identified in the *Aesthetics* theme that were related to the online content somehow failing to meet the expectations of the participants. In this theme, however, some of those cues were more ambiguous than others and evoked more general sentiments of unease with the content, rather than identification of specific cues that provoked concern. For example, there were 18 instances of participants simply commenting that an image was "generic looking" and 26 comments to the effect that something in the image was "weird."

On the other hand, some comments were much more specific and directly related to the online content's appearance. The comments below demonstrate this specificity as it relates to text, typography, and formatting.

- "Text indistinct, not consistent with other text and appears out of place"
- "This doesn't match AT&T's typography"
- "This is all horrible. It is trying to mimic an office365 login form with a google form and the formatting and everything look horrible as a result."

Comments in this theme also targeted the use of templates and free hosting services as suspicious as well, which were cues that had not previously been identified in the taxonomy of phishing cues. Examples of these comments include, "Low-res 'header' image clearly done in Google Forms" and "suggests page was built using a free website builder."

V. DISCUSSION

There were a few conceptually related categories within each theme where similar cues were independently identified, indicating an emergent theme, Failure to Meet Expectations. While most phishing cues can be interpreted as some sort of failure to meet expectations, this theme specifically encompasses those categories where participants called out the mismatch between expectations and execution, in contrast to other categories and themes.

For example, cues about the use of platform templates and free hosting, such as Weebly, WordPress, GoDaddy, etc. or the use of generic form and document templates like Google Forms and Google Docs were prevalent in each theme and spoke to the unlikeliness of their use in certain situations. In each theme, use of these tools was suspicious and worthy of comment either in general terms such as, "Google forms" reported in Aesthetics and "This brand would not use a Google Form" in Previous Interactions, to questioning why the form was used in Interpersonal Trust, "Why would AT&T have a google forms login page?". Similarly, the use of a free service by a site purporting to represent a major telecom company defied expectations and failed to convince participants of the content's legitimacy.

As mentioned in the introduction, anti-phishing training tools require content that is recent and labeled accurately. Our prior analyses indicated that crowdsourcing can provide accurate labeling [12], and the current study demonstrates that crowd-generated labels can help identify a more comprehensive range of cues for inclusion in training. This addresses the content recency requirement in two ways: first, the crowdsourced strategy is a scalable approach for regularly annotating new phishing content so that it stays fresh and relevant, and second, the crowd has the ability to identify novel phishing cues. In each case, our research shows that crowdsourcing can help pinpoint signals of new phishing strategies as they emerge while also providing a more robust set of labels to meet content development requirements for training purposes.

To operationalize crowdsourcing for addressing content recency, software architects would need to a) build a pipeline for sourcing phishing content from online repositories, b) integrate with one or more crowdsourcing platforms to allow

crowd users to markup the content, c) apply crowd label deconfliction techniques such as [19] to ensure the crowd reaches a consensus, and then d) import the crowd-labeled content into their tool's content library. These steps, while not trivial or cost-free, are significantly less costly than employing teams of cybersecurity experts in an ongoing content development pipeline. Crowd users provided their services for free in our work (as it was a research project), and would work on paid platforms at significantly reduced cost compared to experts, meaning that most of the cost burden is up-front in the additional development required to integrate with crowd tooling providers. Another benefit of this strategy is the speed of labeling: the corpus that required months for our team to assemble was exhaustively labeled in under a week. Once an effective crowd content development platform is established, it could also be modularized for use across other training tools.

Overall, we saw that crowdsourced content labeling provides multiple advantages for the content requirements of anti-phishing training tools. Although the results suggest that crowdsourcing can support the labeling step in the process of content generation, the upstream process of capturing phishing (and legitimate) content for labeling and subsequent use remains a challenge, and more work is needed to resolve this bottleneck. In addition, further analysis of the non-specific comments mentioned above would help to identify commonalities among the latent cues that participants intuitively responded to while being unable to articulate the particular details that appeared suspicious.

A. Limitations and Future Work

As with all proof-of-concept research, this study has several limitations. Content curated for the labeling task was constrained by our ability to obtain phishing material from existing repositories and capture an adequate range of complementary legitimate content that spans a variety of contexts that often appear in phishing. We utilized multiple resources for phishing content, which required extensive manual effort to eliminate duplicated and "NSFW" phishing messages, with legitimate content from both primarily North American and UK sources, as most Zooniverse volunteers are from these regions and cultural context is a relevant consideration for labelers' ability to identify certain types of cues. A related limitation is that the people who are willing and able to participate in online volunteer labeling projects may be less diverse than the population of potential users for anti-phishing training tools, and is certainly less diverse than the global population that is subject to phishing attempts. However, we believe that crowdsourcing is a step in the right direction, as it draws on an inherently greater diversity of perspectives than strategies that rely on labeling from cybersecurity experts.

Further, our methods relied on crowd labeling, which is essentially an inductive content analysis strategy, and meaningful cues labeled with themes that appeared very infrequently may have been overlooked. For this analysis, we did not evaluate consensus on 'Other Phishy Findings' at the cue level, instead focusing on categories that were raised repeatedly by multiple labelers. In some cases, the same cue in the same image was commented upon by several people, while in others, the same cues were identified in multiple images but with fewer individuals flagging each instance. A more granular analysis of these data will require building on the work presented here, creating an opportunity for future research.

Finally, we focused our analysis on just three themes that our research team had applied to a large number of 'Other Phishy Findings' comments, but other themes in the taxonomy would benefit from similar scrutiny in future work. Per the Zooniverse researcher agreement, we will release the data set upon publication of our full complement of primary results, enabling others to tackle these research challenges and explore this unique data set.

VI. Conclusion

In recent years, much research has been dedicated to automated techniques for phishing detection, with less being dedicated to phishing awareness training development. However, little attention has been given to the experiences of participants on the receiving end of phishing attempts and the indicators that stand out to them as signals of phishy content. In this paper, we looked at how the wisdom of the crowd could be applied to obtain multiple perspectives on current phishing techniques, with crowdsourced data collected from an observational study on Zooniverse, to help identify what cues were understood to be phishy by a crowdsourced audience. In particular, our analysis of the primary categories Previous Interactions, Engagement-Cognitive-Aesthetics, and Interpersonal Trust-Organizational identified failed expectations, typography, and bad formatting as the most commonly mentioned cues in volunteers' comments. The results support further exploration of crowdsourcing to address the ongoing content labeling requirements for anti-phishing training tools.

Acknowledgment

We thank the thousands of Zooniverse volunteers who contributed to Phish Finders, and Jude Lowe and Keegan Shanahan for contributing to the analysis. This work was supported in part by a grant from the University of Nebraska at Omaha's University Committee on Research and Creative Activity.

References

[1] M. Steves, K. Greene, and M. Theofanos, "Categorizing human phishing difficulty: a phish scale," *Journal of Cybersecurity*, vol. 6, no. 1, p. tyaa009, 2020.

[2] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: a literature survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091–2121, 2013.

[3] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 581–590.

[4] KnowBe4. Security awareness training | KnowBe4. [Online]. Available: https://www.knowbe4.com

[5] CyberTrain. [Online]. Available: https://www.cybertraininc.com

[6] J. Staggs, R. Beyer, M. Mol, M. Fisher, B. Brummel, and J. Hale, "A perceptual taxonomy of contextual cues for cyber trust," in *Journal of The Colloquium for Information Systems Security Education*, vol. 2, no. 1, 2014, pp. 10–10.

[7] S. Palka and D. McCoy, "Dynamic phishing content using generative grammars," in *2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, Apr 2015, p. 1–8.
[8] C. Steger, B. Butt, and M. B. Hooten, "Safari science: assessing the reliability of citizen science data for wildlife surveys," *Journal of Applied Ecology*, vol. 54, no. 6, p. 2053–2062, 2017. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.12921
[9] A. Matsunaga, A. Mast, and J. A. Fortes, "Reaching consensus in crowdsourced transcription of biocollections information," in *2014 IEEE 10th International Conference on e-Science*, vol. 1, Oct 2014, p. 57–64.
[10] C. Nguyen, M. L. Jensen, A. Durcikova, and R. T. Wright, "A compari-son of features in a crowdsourced phishing warning system," *Information Systems Journal*, vol. 31, no. 3, pp. 473–513, 2021.
[11] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie," p. 9, 2002.
[12] V. K. Ahuja, H. Rosser, A. Grover, and M. Hale, "Phish finders: Improving cybersecurity training tools using citizen science," in *ICIS 2022 Proceedings*, Manuscript submitted for review.
[13] M. L. Hale, R. Gamble, J. Hale, M. Haney, J. Lin, and C. Walter, "Measuring the potential for victimization in malicious content," in *2015 IEEE International Conference on Web Services*, Jun 2015, p. 305–312.
[14] A. Smith, S. Lynn, C. Lintott, and R. Simpson, "Zooniverse-web scale citizen science with people and machines." in *AGU Fall Meeting Abstracts*, 2013.
[15] K. A. Neuendorf, *The Content Analysis Guidebook*. 2455 Teller Road, Thousand Oaks California 91320: SAGE Publications, Inc, 2017. [Online]. Available: https://methods.sagepub.com/book/the-content-analysis-guidebook-2e
[16] M. Hale, C. Walter, J. Lin, and R. Gamble, "A priori prediction of phishing victimization based on structural content factors," vol. 5.
[17] G. Wethor, "Investigating the impact of user interface aesthetic quality on phishing victimization," ISBN: 9780438260726. [Online]. Available: https://www.proquest.com/docview/2089996489/abstract/29218D6A1F2F4D2CPQ/1
[18] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI '90*. Seattle, Washington, United States: ACM Press, 1990, p. 249–256. [Online]. Available: http://portal.acm.org/citation.cfm?doid=97243.97281
[19] I. S. Rosenthal, J. E. Byrnes, K. C. Cavanaugh, T. W. Bell, B. Harder, A. J. Haupt, A. T. Rassweiler, A. Pérez-Matus, J. Assis, A. Swanson *et al.*, "Floating forests: Quantitative validation of citizen science data generated from consensus classifications," *arXiv preprint arXiv:1801.08522*, 2018.