datasets (*e.g.* [36]). More often, links or request information for the datasets are buried in some footnote or a specific section on manuscripts, making it challenging to discover. Half-way through our dataset collection, we were excited to see Google deploy the new Dataset Search engine [37]. Unfortunately, at that time only 1 dataset related to accessibility, VizWiz [7] surfaced. To promote discovery and transparency for accessibility datasets, we use our initial dataset collection to pre-populate and deploy IncluSet [18], an accessibility dataset repository that only stores metadata linking to the data source and description while supporting data discovery through the Google Schema [38].

Out of the 140 datasets that were manually located over a two-year period, only 56 can be downloaded directly (*e.g.*, through a webpage from the dataset creators) and 31 are available upon request (*e.g.*, a given email by the creators). The remaining 53 don't include any sharing intent or information (we still link to them as they fit our criteria). This is not a surprise. The majority of human-computer interaction researchers that work with these populations do not share data. When looking at 509 papers on wellness, accessibility, and aging published at ACM CHI 2010–2018, Abbot *et al.* [17] found that only 3 made their data publicly available. This number is quite low when compared to prior work surveying CHI authors from the same period [39]; researchers found that out of 373 reporting or generating any type of data, 80 shared raw data. Reasons for not sharing included data sensitivity, participant consent, and re-identification risks. This difference could be explained by increased privacy risks for accessibility data that are amplified by the risk of disability disclosure. In our collection, we see that this non-sharing strategy is not unique to a specific population; it is prevalent across data from different user groups including those with visual, hearing, cognitive, speech, and mobility impairments as well as autism. More so, we observe that all of our datasets from people with developmental impairments follow this strategy. Another observation is that children are often involved in these unshared datasets (*e.g.*, [40, 41]). The only publicly available dataset collected from children in our repository included eye tracking measurements of autistic children [42].

3.1 Publicly Available Accessibility Datasets

Datasets in this group can be directly downloaded from personal and project-specific websites (36); repositories like Kaggle (4), UCI Machine Learning Repository (3), and PhysioNet (5); OrtoLang (3); Zenodo (3); Synapse.org (1); and Open Science Foundation (1). This strategy was most commonly found across datasets from people who are deaf/Deaf or hard-of-hearing (25), typically including sign language videos and gloss annotations. The majority of them were shared by computational linguists and computer vision researchers. We also see this sharing strategy for data from people with motor impairments (14) *e.g.*, providing touchscreen gestures for users with upper body motor impairments [43]. The majority of these datasets fall under both motor and cognitive categories (9) as they typically involve people with progressive conditions such as Parkinson's and Huntington where symptoms relate to motor and cognitive abilities. Though the motivation for collecting these datasets may differ, there is an underlying potential for "detecting" such conditions. This strategy of direct download was also common among datasets sourced from people who are blind or have low vision (7) sharing their photos (*e.g.*, [44]), touchscreen gestures (*e.g.*, [45]), and walking patterns (*e.g.*, [46]).

Sharing License. The majority of publicly available datasets did not provide any license information, form of agreement, or requirement for downloading (31 out of 56). Those who did, mainly opted for the creative commons family of licenses (CC: 11, CC BY-NC 2.0: 3, CC4.0: 1). Few chose ODC Public Domain Dedication and License (5) and New BSD License (2). One dataset used a custom license such as *Synapse Commons Governance* and one declared the data under a specific copyright holder but did not provide a license.

Anonymization/Privacy. We highlight two more recent efforts found among these datasets that consider re-identification and privacy risks. The first one relates to detecting progression of Parkinson's disease, where researchers make a conscious decision to use only non-speech sounds like breathing, clearing throat and swallowing to predict the risk of onset [47]. However, the risk for disability disclosure remains. The second one, relates to visual question answering systems, where researchers attempt to recognize the presence of private information in images taken by blind individuals [44].

Funding. The majority of the datasets indicate support in their acknowledgments from public funding such as NSF and the European Union. We observe that sharing efforts start around year 2000 for the health population and populations related to vision, mobility, hearing, and cognitive impairments; after 2010 for datasets related to speech impairments; and after 2015 for autism. We haven't found any publicly available datasets sourced from people with developmental or learning impairments.

3.2 Datasets Shared Upon Request

Datasets in this group can be accessed only upon request through specific procedures. The most common practice we observe is to have a dedicated dataset webpage with a note to contact one of the authors (typically the Project Investigator) given an email address without any further details on eligibility or process. Another practice is to describe the license agreement, the requirements to obtain the data, as well as the types of data that would be shared. This information was either included on the project webpage or included on a dedicated section of the publication where the data were introduced, named *Distribution*. For example, in the BosphorusSign dataset [48] this section reads: *"The collected corpus will be available to download for academic purposes upon filling a license agreement available from the BosphorusSign website. The provided data will include ... "*

A contrasting pattern across many of the populations sourced in datasets shared upon request is that they fall under what is called "invisible disabilities," disabilities that are less apparent to others and perhaps more sensitive for disclosure. For example, we see here datasets from people with language and learning impairments, which were not publicly shared. Overall, this strategy was most often adopted for datasets sourced from people with cognitive impairments (16) such as people with dementia, Alzheimer's disease, or people with mental-health issues. Motivated by early diagnosis or detection they include logs of daily activities, in video or audio formats, or interaction events with computing devices. Though less in number than the publicly available datasets, we see here datasets generated by people who are deaf/Deaf or hard-of-hearing (9) focusing mostly on analysis of linguistic phenomena that can contribute to sign language synthesis (*e.g.*, [20]) and recognition (*e.g.*, [48]). This sharing strategy seems also to be more prominent for datasets sourced from people with speech impairments (7) with a goal to improve speech recognition (*e.g.*, dysarthric speech [49]) and assessment tools (*e.g.*, [50]).

Sharing License. Almost all datasets available upon request did not provide any license information, form of agreement, or conditions for access prior to contacting (28 out of 31). Few exceptions included the DEVISIGN datasets [51], which detailed a procedure and specified an application format, and the dataset by Avgerinakis *et al.* [52], which mentions non-commercial usage with additional information to be revealed upon request.

Anonymization/Privacy. We highlight in chronological order two efforts found among these datasets that consider re-identification and privacy risks. The first one relates to dementia detection through videos of activities, where researchers prohibit those requesting the dataset from linking individual data to any other information, prevent them from contacting any participant, and forbid the use of participants' face in publications of any kind [52]. The second one, relates to sign language corpora, where is difficult to hide the visual appearance of the signers as facial expressions and head movements are critical for conveying meaning. Here, researchers attempt to anonymize name entities by making relevant signs or mouthing components unrecognizable [53].

Funding. All datasets indicate support from public funding (29) (*e.g.*, NSF and the European Union) and industry (2) (*e.g.*, Microsoft). We observe that one of the first sharing efforts was in 1984 for language and cognitive impairments populations; around 1995 for speech impairments and health populations; and 2010-2016 for hearing, mobility, learning, and vision impairments as well as autism. We haven't found any datasets available upon request from people with developmental impairments.

4 Conclusion

Datasets directly sourced from underrepresented communities such as people with disabilities and older adults can contribute to more inclusive AI applications as well as innovative assistive technologies. However, they are scarce. We discuss challenges for locating such datasets and provide a data surfacing repository to help with their discovery. More so, we present unique challenges and privacy risks for collecting and sharing these datasets and discuss how strategies (shared publicly, shared upon request, and unshared) across the 140 datasets prepopulating our repository, differ across populations and research communities. We find that beyond lack of standardization, the majority of shared datasets lacked any license information, form of agreement, or condition for access. Also very few of them address potential re-identification and privacy risks. We call for better sharing practices as well as technical, legal, and institutional privacy frameworks that are more attuned to concerns from these communities *e.g.*, risks of inaccurate or non-consenting disclosure of a disability.

Broader Impact

As machine learning expands its role in decision making processes, so does the impact of the underrepresented training and benchmarking data for the life of people with disabilities affecting their employment, economic self sufficiency, independence, inclusion and integration into society. Given the increasing attention in machine learning to concerns of fairness and ethics, we have an opportunity to ensure that people with disabilities and other underrepresented communities involved in wellness, accessibility, and aging are part of this conversation. In this paper we discuss why datasets directly sourced by these communities are scarce with a focus on data sharing risks. We hope that the insights from our analysis of sharing practices across 140 datasets from 1984 to 2019 to inform and motivate appropriate curation and use of such datasets. More so, to promote research and educational efforts that can benefit these communities we have deployed a data surfacing repository for accessibility datasets. We note that our repository is not a call to include underrepresented communities, that we aim to benefit, in models that follow rigid categorization that can pose risks for non voluntary disability disclosure. On the contrary, we are hoping it will help us better understand sharing practices and potential concerns that can feed into the conversations to follow.

Acknowledgments and Disclosure of Funding

We thank our anonymous reviewers for their insightful feedback on an earlier version of this paper. This work is supported by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR), ACL, HHS (#90REGE0008). The opinions herein are those of the authors.

References

[1] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 2124–2132. AAAI Press, 2017.

[2] Benedikt Fecher, Sascha Friesike, and Marcel Hebing. What drives academic data sharing? volume 10, pages 1–25. Public Library of Science, 02 2015.

[3] Ingeborg Meijer, Stephane Berghmans, Helena Cousijn, Clifford Tatum, Gemma Deakin, Andrew Plume, Alex Rushforth, Adrian Mulligan, Sarah de Rijcke, Stacey Tobin, Thed Van Leeuwen, and Ludo Waltman. Open data: the researcher perspective. CWTS, Universiteit Leiden, Leiden., 04 2017.

[4] Carol C. Diamond, Farzad Mostashari, and Clay Shirky. Collecting and sharing data for population health: A new paradigm. volume 28, pages 454–466, 2009.

[5] Mark Walport and Paul Brest. Sharing research data to improve public health. In *The Lancet*, volume 377, pages 537–539. Elsevier, 2019/09/15 2011.

[6] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition. In *Sixteenth annual conference of the international speech communication association*, pages 478–482, 2015.

[7] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. IEEE, Jun 2018.

[8] Hernisa Kacorri. Teachable machines for accessibility. Number 119, page 10–18, New York, NY, USA, November 2017. Association for Computing Machinery.

[9] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudrealt, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. Sign language recognition, generation, and translation: An interdisciplinary perspective. 2019.

[10] Meredith Ringel Morris. Ai and accessibility. volume 63, page 35–37, New York, NY, USA, May 2020. Association for Computing Machinery.

[11] Andrew Sears and Vicki L. Hanson. Representing users in accessibility research. volume 4, pages 7:1–7:6, New York, NY, USA, March 2012. ACM.

[12] Carol Neidle, Ashwin Thangali, and Stan Sclaroff. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*. Citeseer, 2012.

[13] Hernisa Kacorri. *Data-Driven Synthesis and Evaluation of Syntactic Facial Expressions in American Sign Language Animation*. PhD thesis, CUNY Academic Works, 2016.

[14] Foad Hamidi, Kellie Poneres, Aaron Massey, and Amy Hurst. Who should have access to my pointing data?: Privacy tradeoffs of adaptive assistive technologies. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, pages 203–216, New York, NY, USA, 2018. ACM.

[15] Jutta Treviranus. The value of being different. In *Proceedings of the 16th Web For All 2019 Personalization - Personalizing the Web*, W4A '19, pages 1:1–1:7, New York, NY, USA, 2019. ACM.

[16] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. Toward fairness in ai for people with disabilities: A research roadmap. 2019.

[17] Jacob Abbott, Haley MacLeod, Novia Nurain, Gustave Ekobe, and Sameer Patil. Local standards for anonymization practices in health, wellness, accessibility, and aging research at chi. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery.

[18] Hernisa Kacorri, Utkarsh Dwivedi, Sravya Amancherla, Mayanka Jha, and Riya Chanduka. *IncluSet: A Data Surfacing Repository for Accessibility Datasets*. Association for Computing Machinery, New York, NY, USA, 2020.

[19] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign Language Recognition Using Sub-units. In Sergio Escalera, Isabelle Guyon, and Vassilis Athitsos, editors, *Gesture Recognition*, The Springer Series on Challenges in Machine Learning, pages 89–118. Springer International Publishing, Cham, 2017.

[20] Pengfei Lu and Matt Huenerfauth. Cuny american sign language motion-capture corpus: first release. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, The 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey*, 2012.

[21] R. Vatavu, B. Gheran, and M. D. Schipor. The Impact of Low Vision on Touch-Gesture Articulation on Mobile Devices. volume 17, pages 27–37, January 2018.

[22] Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Hironobu Takagi, and Chieko Asakawa. Supporting orientation of people with visual impairment: Analysis of large scale usage data. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '16, pages 151–159, New York, NY, USA, 2016. ACM.

[23] Ugo Cesari, Giuseppe De Pietro, Elio Marciano, Ciro Niri, Giovanna Sannino, and Laura Verde. A new database of healthy and pathological voices. volume 68, pages 310–321, May 2018.

[24] Luz Rello, Ricardo Baeza-Yates, and Joaquim Llisterri. DysList: An annotated resource of dyslexic errors. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1289–1296, Reykjavik, Iceland, May 2014. European Languages Resources Association (ELRA).

[25] Burn Model System National Data and Statistical Center. Burn model system: Advancing recovery through knowledge, 1994.