A research infrastructure for generating and sharing diversity-aware data

Matteo BUSSO [a,1], Ronald CHENU ABENTE ACOSTA [b], Amalia DE GÖTZEN [b]
[a] *Department of Information Engineering and Computer Science, University of Trento*
[b] *Department of Architecture, Design and Media Technology, Aalborg University*
ORCiD ID: Matteo Busso https://orcid.org/0000-0002-3788-0203, Ronald Chenu Abente Acosta https://orcid.org/0000-0002-1121-0287, Amalia de Götzen https://orcid.org/0000-0001-7214-5856

Abstract. The intensive flow of personal data associated with the trend of computerizing aspects of people's diversity in their daily lives is associated with issues concerning not only people protection and their trust in new technologies, but also bias in the analysis of data and problems in their management and reuse. Faced with a complex problem, the strategies adopted, including technologies and services, often focus on individual aspects, which are difficult to integrate into a broader framework, which can be of effective support for researchers and developers. Therefore, we argue for the development of an end-to-end research infrastructure (RI) that enables trustworthy diversity-aware data within a citizen science community.

Keywords. Diversity, Citizen Science, Research Infrastructure, Data Collection, Data Sharing

1. Introduction: the relevance of a diversity aware-RI

Digitization [1] is exponentially increasing the production of data and driving relevant economic, social, and political changes [2]. The trend is often associated with datafication [3], namely the act of quantifying and computerising people everyday life and that, according to [4], has the potential to shape both the ontologies and the methodologies of many disciplines.

However, as [5] would say, "bigger data are not always better data". Indeed, many useful data to model people everyday life as often missing [6], which leads to the problem of reducing people to their average, ignoring and disincentivizing their diversity, namely how similar or different are a person's experience, competences or traits with regards others [7].

Secondly, a growing body of literature is focusing on bias and bias management (see, e.g., the extensive work done by [8,9,10]) in scientific fields such as AI, health, and behavioural studies.

Finally, the way in which personal data is treated is not risk-free, especially when considering aspects such as consumer privacy, non-transparent legal regulation, or even bias

[1] Corresponding Author: Matteo Busso, PhD Candidate, email: matteo.busso@unitn.it

in the programming. Examples are the processing of data for the purpose of advertising conducted by Google or Facebook [11], but also all the documented cases of "untrustworthy" AI (see e.g., the cases related to face recognition [12]).

Although many strategies to mitigate the risks associated with non-diversity-aware approaches to data have already been proposed, such as explainability [13] or the report on Ethics guidelines for trustworthy AI [14], we believe that, following [5] suggestion, a broader (and radical) approach should be taken.

This is the reason why we suggest the development of an end-to-end research infrastructure (RI)[2] that enables trustworthy diversity-aware data within a citizen science community.

Data management is a complex and multidisciplinary process, ranging from ethical and legal to social sciences and AI. Furthermore, it involves various phases, from collection to preparation up to distribution and reuse (see, e.g., [16]). Furthermore, data is used in numerous fields, both for research and innovation. Being RIs pivotal for developing research areas as they consolidate both technologies and methodologies (considering, for example, standards or guidelines), we believe that an end-to-end RI is necessary, to support the researchers or developers within the whole data management process.

Then, we consider diversity-aware data, in order to represents the uniqueness of people within their context. An analogous term is Big Thick Data, [17], which aims to combine Big (Thin) Data, which are usually high in volume but provides little or no contextual information [6], as can be data coming from smartphone sensors (e.g., GPS location, WiFi connection, app usage); and Thick Data, which are contextual data provided by the interaction with the person. Thick interactions can concern a variety of interplay, e.g., of the person with her context or with other people and of the person with the machine collecting the data. From the first interplay derives characteristics such as gender, age or nationality, but also deeper aspects, such as emotions or values connected to specific events; from the latter, it derives feedback on the machine usages. Collected together, this information allows to map people diversity in all its aspects.

In this sense, diversity-aware data not only allows for the representation of the diversity of people, but it is what enables effective Hybrid Human-Artificial Intelligence approaches, where AI adapts to the human via Thick interactions. In the next section, we will focus on applications that allow the collection of Big Thin and Thick Data via smartphone, which is the pervasive tool par excellence.

Finally, by trustworthy, we mean a structure that not only complies with the guidelines proposed by the EU commission and the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April) but which is able create a relationship of trust with people who provide their data, e.g., through transparent communication, but also acting as an intermediary in defending their interests.

In this sense, it is crucial the relationship between those who provide the data and those who analyse it. Therefore, to foster the diversity-aware approach, a community of trust need to be created. We propose to follow the consolidated approach of Citizen Science (CS) [18], aiming to involve citizens not only in research but also in a shared data culture.

The remainder of the paper is organized as follow. Section 2 describes the current status on diversity-aware data generation and sharing. Section 3 presents a solution out-

[2] According to [15], RIs are "facilities that provide resources and services for the research communities to conduct research and foster innovation in their fields".

line and challenges for developing a diversity-aware following the exemplary case of [19] within the DataScientia ecosystem. Section 4 closes the paper.

2. The current status on diversity-aware data generation and sharing

Although an end-to-end RI addressing diversity-aware data doesn't exist yet, there are several technologies and infrastructures that address parts of it. In particular aspects of (i) data collection; (ii) data management and distribution; (iii) involving people in experiments or within a community.

Data collection Although increasingly fundamental to CS, "sensing technologies [...] is one area that has not yet been harnessed" [20], both from a theoretical and technological point of view. Considering diversity-aware data collection, of the several configurable data collection applications, few are able to collect them. Many applications, such as Psychlog [21] and Mobile Sensing Platform [22], collect data only from user interactions, while others, such as [23], collect only sensors data. Two applications are particularly relevant, namely AWARE [24], which is a complex system often adopted within the ESM framework[3] [25], but it does not provide data management support, and [26], whose ease of configuration makes it suitable for a CS community, even if not equipped for collecting all the sensors data.

Data management and distribution The disciplines that deal with personal data often develops RIs to support researchers in the aspects of data management and sharing. For instance, within social sciences, [27] and [28] provide support for ethics assessment and data management and, alongside with [29], they enable the documentation and distribution of high quality survey data. Data distribution is particularly advanced in the healthcare sector, with leading RIs such as [30] or [31], which also have played a key role in the management of the Covid19 pandemic.

However, despite the obvious support provided by such infrastructures, they are not end-to-end. Furthermore, they remain tied to individual research communities, not favouring effective interdisciplinary exchange.

Crowd-sourcing vs. Communities According to [32], CS has some aspects in common with crowd-sourcing, especially in involving non-expert people in fulfilling research tasks. Examples are participatory sensing [33] and Mobile Crowd Sensing [34] and they are particularly relevant as they rely on the pervasiveness of smart devices to collect data on large panel, even though based on people often coming from Western countries, which is a main issue for considering the diversity of people.
However, crowd-sourcing considers the participant only as a contributor to the data collection [35], but rather than an active member of a community, as it does not considers an involvement in the research process nor education and information projects. On the contrary, projects like [36], [37] or [38] can be considered as actual CS communities, even if their focus is on natural sciences and not on the diversity of people and their behaviour, while [39] and [40] have a broader focus, even though they are not based on an end-to-end RI.

[3] Experience Sampling Method (ESM) is an intensive longitudinal social and psychological research methodology, i.e. designed for reducing social and cognitive bias in data collection, where participants are asked to report on their thoughts and behaviours.

3. Towards a diversity-aware RI

To outline a potential solution, we will present the LivePeople case study and discuss some limitations. Even if not yet fully operational, LivePeople contains a set of proposals of services and technologies that covers the main aspects described in Section 2 for creating an end-to-end diversity-aware RI embedded in a CS community.

Data collection One of the LivePeople services is a cutting-edge data collection app called iLog [41][4], which allows collecting diversity-aware information through the interaction with the person and from all the smartphone sensors.

Data management and distribution The whole data management process is ethics and privacy-aware by design, and it is based on a consolidated methodology considering quality standards from the social science domain [16], and following the [46]. Regarding this latter principle, the RI also focuses on advanced data integration approaches [47], which aims to extend the collected data for interdisciplinary reuse.

The CS community LivePeople will be established on a cross-country panel of people, i.e. it will be based on the diversity of people. To ensure trust in the community, people will remain the owners of their data and have the option to donate or sell it in exchange for research and services of interest to them. Ultimately, RI will be community-based and community-led. Not only will the RI provide services to community members in their context, but the community itself will be self-sufficient to create and run new projects and to support and contribute to existing ones.

Limits Even if part of the RI has already been applied in different projects (e.g., [48,49]), leading to interdisciplinary publications (e.g., [43,50]), some key aspects of LivePeople are not yet consolidated or validated. Examples are (i) the usability of the iLog app by non-expert users, such as citizens; (ii) the validation of data management outcomes to effective reuse of resources - both in terms of data quality and their interoperability; (iii) the lack of a panel that can be consulted on demand and incentive strategy that guarantees high collaboration from members within the community.

4. Conclusion

In this paper we argued how datafication is affecting data management and data quality creating bias in their reuse. Then we showed how the current status of diversity-aware data generation and sharing platform are not suitable for the purpose of creating trust and quality data, and we presented a former solution, considering some of its constrains.

Acknowledgments

The work is funded by the *"WeNet - The Internet of Us"* Project, funded by the European Union (EU) Horizon 2020 programme under GA number 823783.

[4] [42,43,44,45] is a list of publications which describe the use of iLog and of iLog collected data in various studies.

References

[1] Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[2] Claudia Loebbecke and Arnold Picot. Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems*, 24(3):149–157, 2015.

[3] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[4] David M Berry. The computational turn: Thinking about the digital humanities. *Culture machine*, 12, 2011.

[5] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[6] Nigel G Fielding, Raymond M Lee, and Grant Blank. *The SAGE handbook of online research methods*. Sage, 2008.

[7] Laura Schelenz, Ivano Bison, Matteo Busso, Amalia De Götzen, Daniel Gatica-Perez, Fausto Giunchiglia, Lakmal Meegahapola, and Salvador Ruiz-Correa. The theory, practice, and ethical challenges of designing a diversity-aware platform for social relations. In *AIES'21*. Association for Computing Machinery, 2021.

[8] Jahna Otterbacher. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1955–1964, 2015.

[9] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, pages 933–936, 2018.

[10] Kalia Orphanou, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Fausto Giunchiglia, Veronika Bogina, Avital Shulner Tal, Alan Hartman, and Tsvi Kuflik. Mitigating bias in algorithmic systems—a fish-eye view. *ACM Computing Surveys*, 55(5):1–37, 2022.

[11] Asunción Esteve. The business of personal data: Google, facebook, and privacy issues in the eu and the usa. *International Data Privacy Law*, 7(1):36–47, 2017.

[12] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[13] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer, 2019.

[14] EUCommission. Ethics guidelines for trustworthy ai. `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`, 2023. Accessed: 2023-05-07.

[15] EUCommission. Research and innovation. `https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/research-infrastructures_en`, 2023. Accessed: 2023-05-07.

[16] Louise Corti, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard. *Managing and sharing research data: A guide to good practice*. Sage, 2019.

[17] Tobias Bornakke and Brian L Due. Big–thick blending: A method for mixing analytical insights from big and thick data sources. *Big Data & Society*, 5(1):2053951718765026, 2018.

[18] Mordechai Muki Haklay, Daniel Dörler, Florian Heigl, Marina Manzoni, Susanne Hecker, Katrin Vohland, et al. What is citizen science? the challenges of definition. *The science of citizen science*, 13, 2021.

[19] LivePeople. `https://datascientiafoundation.github.io/LivePeople/`, 2023. Accessed: 2023-05-07.

[20] Michael J O'Grady, Conor Muldoon, Dominic Carr, Jie Wan, Barnard Kroon, and Gregory MP O'Hare. Intelligent sensing for citizen science: challenges and future directions. *Mobile Networks and Applications*, 21:375–385, 2016.

[21] Andrea Gaggioli, Giovanni Pioggia, Gennaro Tartarisco, Giovanni Baldus, Daniele Corda, Pietro Cipresso, and Giuseppe Riva. A mobile data collection platform for mental health research. *Personal and Ubiquitous Computing*, 17:241–251, 2013.

[22] Skyler Place, Danielle Blanch-Hartigan, Channah Rubin, Cristina Gorrostieta, Caroline Mead, John Kane, Brian P Marx, Joshua Feast, Thilo Deckersbach, Alex "Sandy" Pentland, et al. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *Journal of medical Internet research*, 19(3):e75, 2017.

[23] ResearchStack. `http://researchstack.org/`, 2023. Accessed: 2023-05-07.

[24] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[25] Mihaly Csikszentmihalyi, Mihaly Csikszentmihalyi, and Reed Larson. Validity and reliability of the experience-sampling method. *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi*, pages 35–54, 2014.

[26] CitizenScienceLogger. `https://lab.citizenscience.ch/en/tools/cslogger`, 2023. Accessed: 2023-05-07.

[27] UkDataArchive. `https://www.data-archive.ac.uk/`, 2023. Accessed: 2023-05-07.

[28] GESIS. `https://www.gesis.org/en/home`, 2023. Accessed: 2023-05-07.

[29] Open Science Framework (OSF). `https://osf.io/`, 2023. Accessed: 2023-05-07.

[30] ComputerOntario. `https://www.computeontario.ca/what-is-dri`, 2023. Accessed: 2023-05-07.

[31] BioBank. `https://www.biobank.it/`, 2023. Accessed: 2023-05-07.

[32] Mahmood Hosseini, Alimohammad Shahri, Keith Phalp, Jacqui Taylor, and Raian Ali. Crowdsourcing: A taxonomy and systematic mapping study. *Computer Science Review*, 17:43–69, 2015.

[33] Jeffrey Goldman, Katie Shilton, Jeff Burke, Deborah Estrin, Mark Hansen, Nithya Ramanathan, Sasank Reddy, Vids Samanta, Mani Srivastava, and Ruth West. Participatory sensing: A citizen-powered approach to illuminating the patterns that shape our world. *Foresight & Governance Project, White Paper*, pages 1–15, 2009.

[34] Huadong Ma, Dong Zhao, and Peiyan Yuan. Opportunities in mobile crowd sensing. *IEEE Communications Magazine*, 52(8):29–35, 2014.

[35] Jennifer L Shirk, Heidi L Ballard, Candie C Wilderman, Tina Phillips, Andrea Wiggins, Rebecca Jordan, Ellen McCallie, Matthew Minarchek, Bruce V Lewenstein, Marianne E Krasny, et al. Public participation in scientific research: a framework for deliberate design. *Ecology and society*, 17(2), 2012.

[36] Zooniverse. `https://www.zooniverse.org/`, 2023. Accessed: 2023-05-07.

[37] iNaturalist. `https://www.inaturalist.org/`, 2023. Accessed: 2023-05-07.

[38] CornellLab. `https://www.birds.cornell.edu/home/`, 2023. Accessed: 2023-05-07.

[39] SciStarter. `https://scistarter.org/`, 2023. Accessed: 2023-05-07.

[40] EU-citizen.science. `https://eu-citizen.science/`, 2023. Accessed: 2023-05-07.

[41] Mattia Zeni, Ivano Bison, Britta Gauckler, Fernando Reis, and Fausto Giunchiglia. Improving time use measurement with personal big collection - the experience of the european big data hackathon 2019. *Journal of Official Statistics*, 2020.

[42] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. Multi-device activity logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 299–302, 2014.

[43] Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. Mobile social media and academic performance. In *International conference on social informatics*, pages 3–13. Springer, Cham, 2017.

[44] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. Human-like context sensing for robot surveillance. *International Journal of Semantic Computing*, 12(01):129–148, 2017.

[45] Fausto Giunchiglia, Mattia Zeni, and Enrico Big. Personal context recognition via reliable human-machine collaboration. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 379–384. IEEE, 2018.

[46] F.A.I.R.Principles. `https://www.go-fair.org/fair-principles/`, 2023. Accessed: 2023-05-07.

[47] Fausto Giunchiglia, Simone Bocca, Mattia Fumagalli, Mayukh Bagchi, and Alessio Zamboni. itelos-building reusable knowledge graphs. *arXiv preprint arXiv:2105.09418*, 2021.

[48] SmartUniversity. `http://su.disi.unitn.it/`, 2023. Accessed: 2023-05-07.

[49] WeNet. `https://www.internetofus.eu/`, 2023. Accessed: 2023-05-07.

[50] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, José Luis Zarza, Alethia Hume, Luca Cernuzzi, Ivano Bison, Marcelo Dario Ro-

das Britez, Matteo Busso, Ronald Chenu-Abente, Fausto Giunchiglia, and Daniel Gatica-Perez. Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.