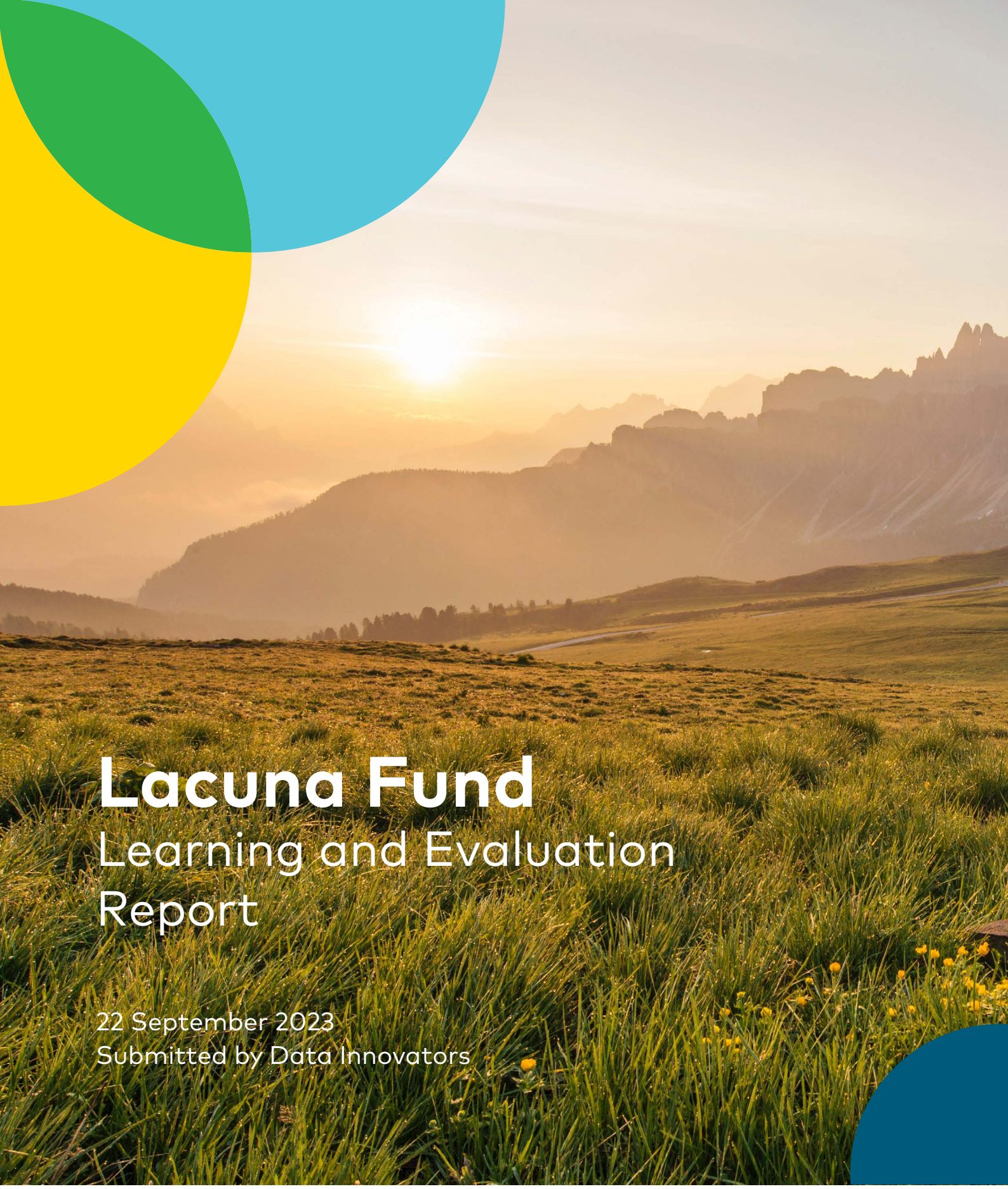


Lacuna Fund

Learning and Evaluation Report

22 September 2023

Submitted by Data Innovators

data INNOVATORS

Lacuna Fund

Our voice in data

Evaluators and authors – Data Innovators
Mutsa Chinyamakobvu
Jerusha Govender
Unathi James
Timothy Leslie
Teboho Letuka
Amanda Ncube

Oversight and contributors – Meridian Institute
Katrina Gehman
Emma Heth
Jennifer Pratt Miles

Key contact
Katrina Gehman - kgehman@merid.org

About Lacuna Fund
Lacuna Fund was established in July 2020 through a collaborative effort between The Rockefeller Foundation, Google.org, Canada's IDRC, and GIZ, representing the German Federal Ministry for Economic Cooperation and Development (BMZ). In 2021, Wellcome Trust, Gordon and Betty Moore Foundation, Patric J McGovern Foundation and Robert Wood Johnson Foundation came onboard and added to the pool of funds to invest in the creation, expansion, and maintenance of datasets used for training or evaluation of machine learning models in low and middle-income contexts.

Acknowledgments
This evaluation was sponsored by Canada's International Development Research Centre (IDRC). We greatly appreciate the time and insights shared by grantees, bellwethers including funders, Technical Advisory Panel members and field experts, and Meridian Institute staff during interviews and sense-making sessions. The inputs of these stakeholders have contributed to comprehensive documentation of lessons and recommendations for the continuous improvement of Lacuna Fund.

Images used in this report:

Page 6: Photo taken near Bungoma in Western Kenya of a farmer with insured crops, including beans and maize. This Lacuna Fund-supported project, "Eyes on the Ground Image Data" consists of a large machine learning dataset for crop phenology monitoring of smallholder farmers' fields. Smart phone cameras were used for the georeferenced and time-stamped crop images, following picture-based insurance protocol. This photo was provided by the project team, with credits to AGRIFIN.

All other photos in report (except for the cover): Agriculture, Language, and Equity & Health grantees gather for the second annual Grantee Convening in Kigali, Rwanda, in June 2023 to discuss their datasets and potential use cases. Photo credit: Kigali Forward.

Cover image: Photo by Ales Krivec on Unsplash

Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BMZ	Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung (Federal Ministry for Economic Cooperation and Development)
C&E	Climate and Energy
C&H	Climate and Health
CAPSI	Center for Artificial Intelligence and Public Sector Innovation
E&H	Equity and Health
EPIC Africa	Emerging Public Interest Computing Africa
FAO	Food and Agriculture Organization
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit (German Federal Ministry for Economic Cooperation and Development (BMZ))
LF	Lacuna Fund
LMICs	Low and Middle-Income Countries
ML	Machine Learning
MLHC	Machine Learning for Healthcare
NASA	National Aeronautics and Space Administration
NGO	Non-Governmental Organization
NLP	Natural Language Processing
RFP	Request for Proposals
SRMHR	Sexual, Reproductive and Maternal Health and Rights
TAP	Technical Advisory Panel
ToC	Theory of Change
UN	United Nations
UNDP	United Nations Development Programme
USAID	United States Agency for International Development
WHO	World Health Organization

Contents


acre
AFRICA

1. Executive Summary

Since its inception in 2020, Lacuna Fund's mission is to fund the creation, expansion, and maintenance of training and evaluation datasets that enable the robust application of machine learning (ML) tools of high standard for underserved populations globally. Lacuna Fund's goals are to fill gaps and reduce bias in training data used for ML; enable underserved populations to benefit from AI; and to deepen understanding of how to effectively, efficiently, and equitably fund the development of labelled datasets.

The scope of the evaluation covered Lacuna Fund activities and projects that were selected from 2020-2022. Additional funding opportunities in 2023 and beyond were not included, however, Grantee Convenings held in 2023 for the 2020-2022 grantee cohorts have been mentioned. The report aimed to assess what aspects of Lacuna Fund have effectively and efficiently enabled the creation, expansion, and maintenance of representative and unbiased training datasets for ML; examine process challenges experienced by stakeholders; and provide recommendations for improvement.

The evaluation was guided by ten main evaluation questions. Mixed methods and a utilization-focused approach were used. Data collection comprised of sense-making workshops with 11 Lacuna Fund representatives, 20 grantee interviews and 21 grantee surveys, 13 bellwether interviews of thought leaders in the machine learning (ML) landscape, and 48 dataset user surveys. Furthermore, 59 grantee records of accepted proposals and 130 documents were reviewed and analyzed, and a participatory mapping session was conducted with 17 participating grantees at Lacuna Fund's Grantee Convening in June 2023.

The questions posed in the evaluation focused on addressing Lacuna Fund's processes, outcomes, its influence in the broader ecosystem, and the landscape Lacuna Fund operates in, as shown below.

Table 1. Areas of Inquiry

There are several main insights in this evaluation, namely that Lacuna Fund has supported underserved researchers and data scientists to contribute towards the ecosystem of datasets in low and middle-income contexts; the development of datasets in high impact domains was prioritized; a grant funding model with a principles approach was implemented, Lacuna Fund has increased the availability of unbiased datasets and supported use case development, and researchers benefitted in a wide variety of ways from Lacuna Fund.

Supporting underserved researchers and data scientist to influence the ecosystem of datasets

Lacuna Fund has demonstrated its support for underserved researchers and data scientists to make an impact on the ecosystem of datasets through its grant funding processes.
Lacuna Fund has invested approximately $9.4 million in 59 projects across 54 countries during 2020-2022 through a funding collaborative between The Rockefeller Foundation, Google.org, Canada's International Development Research Centre (IDRC), German development agency GIZ on behalf of the Federal Ministry for Economic Cooperation and Development (BMZ), Gordon and Betty Moore Foundation, Patrick J McGovern Foundation, Robert Wood Johnson Foundation and Wellcome Trust. This pool of funds invested has continued to be distributed to grantees in 2023. Of the 54 countries, Uganda, Kenya, Nigeria and Tanzania hosted the most projects – 14, 12, 11 and 10 projects, respectively. 55% percent of funds awarded between 2020-2022 were granted to researchers in the most popular domains of Agriculture and NLP. In terms of gender representation, 20 out of 59 (34%) lead grantees are female. The grantee gender ratio seven females to nine males in 2022 is an improvement to prior years.

Prioritizing high impact domains

Lacuna Fund has prioritized funding for developing datasets in high impact domains. In 2020 and 2021, initiatives based in the domains of NLP and Agriculture were funded through RFPs that were primarily Africa-focused calls. Equity and Health was funded in 2021, while Climate and Health, and Climate and Energy were funded in 2022. High impact domains are those in which the integration of ML techniques has the potential to bring about meaningful and transformative change by solving complex problems, leading to substantial benefits for individuals, industries, or society as a whole. For example, the portfolio of Natural Language Processing (NLP) researchers supported by Lacuna Found resulted in the creation of NLP datasets that address specific challenges in Africa and other LMICs. This support from Lacuna Fund has laid the foundation to deepen future work in NLP. The processes and systems already developed will be a starting point for future grantees. NLP is important because it creates openly accessible text and speech resources
that enable operations in diverse languages across under served communities globally.

Support for the Agriculture domain was aligned with addressing food security challenges, especially food production. The Health domain strives for high quality and universal health services. The Equity and Health domain focused on building datasets for AI applications that address context specific health challenges in low-resourced settings, whilst the Climate and Health domain focused on addressing the intersection between community population health, climate, and weather.

Building a principle-based grant funding model

Lacuna Fund has created a grant funding model that is principles-based. This is evidenced by the incorporation of accessibility, equity, ethics,

participatory, quality, and impact as principles at the core of Lacuna Fund's work. These principles are used by Lacuna Fund to understand what works and what could be improved. The principle of accessibility was realized through access to funding and resources, improving grantee capacity and skills, and the publication of datasets via open access. Lacuna Fund made a contribution towards addressing inequity in the AI landscape through funding the development of datasets in key domains by- and for under-represented communities. When groups of voices are missing from ML datasets, the resulting models can be biased or completely inaccurate, so Lacuna Fund addresses inequity by funding the the inclusion of under-represented populations and geographies in ML datasets. When marginalized voices are included, the resulting models and applications are more accurate, less biased, and enable greater access to the benefits of AI. Fairness was prioritized in the call for proposals and the communication of decisions. The principle of transformational impact is yet to be realised as more datasets approach completion and more use cases get developed.

Whilst Lacuna Fund's principles are broadly aligned within the decision-making structures and funding processes, particularly in the Technical Advisory Panel (TAP) and the guidance they receive to evaluate proposals, there is room for improvement in measuring and realizing evidence of how funded projects demonstrate alignment to Lacuna Fund's principles. Clear reporting requirements and processes for grantees post-award, as well as efficient communication and turn-around time beyond the application process are areas of improvement for Lacuna Fund.

Increasing unbiased datasets availability and use case development

As of 18 August 2023, 17 Lacuna Fund projects hosting over 30 datasets have been completed and published since 2020. Datasets are published on Harvard Dataverse, ML Hub, Zenodo, Inalco Bomba Reference Corpus, and GitHub. Datasets have been downloaded a total of 407,500 times thus far and have been cited 19 times. There is some indication that datasets could have high impact. Eleven grantees surveyed have used their datasets for modelling, five of whom are integrating their datasets into real life applications, two in the domain of Health, and three in Agriculture. For example, a grantee project developed datasets for ML to detect and classify diseases of crops crucial for food security, whilst another is using health datasets to train ML models to detect tumours from low quality scans in low resource settings. There are likely more modelling activities going on in the broader group of grantees.

From a general survey of dataset users, there was evidence of one use-case of a grantee dataset – involving the use of a dataset for training a ML model to recognize maize plantations. Two intended use cases were identified from Lacuna Fund published datasets: (1) To use the data to research how to curb pollution through bioremediation.; and (2) to explore the effects of expansion of One Acre Fund on deforestation. Given that the majority of grantees have not completed their datasets to date, and that word of Lacuna Fund datasets is yet to spread more widely, it is expected that more use cases will emerge over time as more grantee datasets become publicly accessible.

Influencing intended benefits to researchers

Lacuna Fund had an effect on the benefits experienced by researchers. Grantees stated that the opportunity provided to them by Lacuna Fund resulted in personal growth, skills development, additional funding, spin-off training and model initiatives, as well as visibility and media coverage. A grantee expressed the

benefits of Lacuna Fund's support, saying:

"[AFRICAI/MIKAI] are sponsoring our AI summer school, which is essentially training people on how to create models on that Lacuna data set. They are supporting us to have students come to Vancouver in Toronto this year and present their models. Lacuna Fund has opened up a floodgate."
- Grantee from the Climate & Health domain

The majority of grantees also reported experiencing project related challenges. Resources and time were commonly reported and followed by funding and capacity. The reported challenges were often context specific and were not anticipated in the project planning phase.

Lessons learned

A range of lessons learned have been generated during the evaluation. These lessons serve to inform decision making for program improvement and can provide valuable guidance for improving future endeavors and ensuring better outcomes.

- Grantees collectively shared that they value collaboration with fellow grantees highly and that they would benefit from a grantee platform that facilitates knowledge sharing, connectedness, collaboration, and learning. Such a platform may reduce existing siloes and foster open communication, interdisciplinarity, resource sharing, adaptability, innovation, and long-term relationships. The Grantee Convenings held by Lacuna Fund in Tunis in 2022, and in Rwanda in June 2023 during the course of the evaluation period confirmed these sentiments as grantees shared experiences and connected for future research endeavours. A result of the most recent convening was the establishment of a Slack channel for grantees to continue to connect.
- Grantees expressed that funding processes could be improved by aligning funding with key milestones and by disseminating information about fund requirements and useful resources more frequently.
- Delays in publishing datasets have been experienced by most grantees due to various project related challenges. It is advised to take this into account by reconsidering the grant period for future funds.
- Developing datasets that yield benefits that are sustained over the long-term is a key concern for grantees. A sustainability plan could be supported by collaborations through a multi-stakeholder sustainability committee. A framework to support grantees from end-to-end and initiate new programs to create societal value and promote community engagement would provide a useful platform for supporting growth, wider data sharing efforts, and model integrations.
- Agriculture and NLP remain key domains for dataset development, in addition to other emerging domains. The majority of grantees interviewed state that their datasets will have multiple use cases, particularly in the domains of Agriculture and NLP. For example, types of use cases in Agriculture include plant disease classification, crop yield prediction, crop type mapping, livestock movement prediction and disease transmission. In NLP, a range of advances such as Automatic Speech Recognition (ASR) models, topic modeling, topic classification, and speech to text technologies have been used create datasets intended to support citizens to consume public information (e.g., via radio), read e-books, and benefit from science communication in their first language, as well as hate and offensive speech recognition for improving safety on social media.
- A range of new trends are emerging that are relevant to Lacuna Fund's context. The increased use of models for predicting future

scenarios is likely to create more demand for data. For example, there is an increase in demand for ethical data for future scenario modelling. However, accessing and sharing data has become challenging for researchers developing datasets. There is a growing need for alternative business models such as tiered licensing, Creative Commons license that prescribes how data is used, and data sharing agreement standards. Generative AI is projected to proliferate, and amidst growing concerns over the safety of AI, grantees and Lacuna Fund could play an influential role in the development of responsible AI.

In conclusion, Lacuna Fund has made significant progress in its aim to enable researchers and data scientists to create, expand, and maintain ML datasets. The evaluation has generated some evidence that suggests that Lacuna Fund has supported the creation of ML datasets that are accessible to and used by developers to achieve benefits relevant to communities currently under-represented in data. It is envisaged that increased awareness of Lacuna Fund, and their continued support for unbiased dataset creation will over time result in greater benefits to marginalized communities currently under-represented in data as more grantee datasets are published publicly.

2. Introduction

2.1 Background

Machine learning has demonstrated substantial promise in global development (De-Arteaga et al. 2018). It is widely recognized that datasets created for ML can provide powerful and innovative solutions for addressing challenges in Agriculture, Language, Health and Climate change, not only in the developed world, but also in low and middle-income contexts[1] (Mupangwa et al., 2020; Oriola and Kotzé, 2020; Ahn et al., 2022; Mutai et al., 2021.; Hengl et al., 2021; Ibrahim et al., 2021).

2.1.1 Domain-related challenges being addressed by artificial intelligence

There are a multitude of challenges that artificial intelligence (AI) has, to varying degrees, addressed. In the health sciences, AI has been used successfully in medical image classification, interpretation, and informing treatment in the fields of radiology, pathology, ophthalmology, and gastroenterology, amongst others. AI systems designed for deep learning are tackling not only diagnosis but also risk prediction and treatment, whilst also addressing other challenges such as a lack of capacity for complex image analysis, processing large volumes of data, time inefficiencies, and reducing human error whilst conducting highly repetitive tasks (Rajpurkar et al. 2022).

In the domain of Agriculture, ML has improved gains by providing rich recommendations and insights about crops to farmers (Meshram et al. 2021). AI demonstrated it can be used to improve the performance, accuracy, cost-effectiveness and flexibility of agriculture operations by addressing commonly experienced challenges such as improper soil treatment, irrigation, plant diseases, pest infestation, low yield, and knowledge gaps between farmers and technology (Eli-Chukwu 2021; Sharma 2021).

AI has also been applied to the field of natural language processing (NLP) due to AI's ability to represent and analyze human language computationally. Advances in NLP technologies include machine translation, text categorization, sentiment analysis, spam filtering, information extraction and information summarization (Khurana et al. 2023). Machine translation can attempt to "leave no language behind" by making online content accessible to non-English language speakers, allowing previously excluded people to benefit from improved access to information (Costa-jussà et al. 2022). NLP techniques using AI have also been shown to improve health efforts by using NLP to: identify at-risk populations through analyzing electronic health records and social media; systemically review scientific literature and unpublished data to identify health interventions and outcomes; as well as to provide knowledge generation and translation in the form of expert level advice to health related questions via chatbots and question answer systems (Baclic et al. 2020).

2.1.2 Poor data quality and limited use of AI in and for low and middle-income contexts

The potential of ML in low and middle-income contexts is largely unrealized (De-Arteaga et al. 2018). Globally there are significant gaps[2] in data for ML and these gaps are considered to be the limiting factor in the development of algorithms and scientific progress (Pallada et al., 2021). It has been recognized that data collection and storage in LMICs is fragmented and is limited by a lack of digitization and

[1]LMICs is commonly the acronym for low- and middle income countries. Lacuna Fund focuses on low and middle-income contexts which may include underserved communities or groups in a high-income country.
[2]Low quality data is considered to be data that is missing, incomplete, inconsistent, inaccurate, duplicated or dated (Gudiva et al. (2017).

associated infrastructure (Ciecierski-Holmes et al., 2022; Wahl et al., 2018). Furthermore, there is a lack of high quality unbiased, labeled datasets for ML (L'heureux et al. 2017; De-Arteaga et al. 2018; Veale and Binns, 2017). Poor quality data can have damaging or catastrophic implications.
It can lead to algorithms that generate biased results, inaccurate predictions, flawed decision making, and model failure (Jain et al. 2020). Moreover, biased datasets can have damaging outcomes for marginalized and vulnerable populations (van Leeuwen et al. 2021; Jain et al. 2020).

Since the performance of ML models is dependent on high quality datasets (Jain et al. 2020; Sarkar, 2021), capturing and maintaining high quality datasets that are representative and unbiased is at the heart of training the models (Gudiva et al. 2017). Delivering unbiased labelled datasets for ML is a prerequisite for the accurate identification of risks, prediction of outcomes, and the realisation of the potential of ML to address contemporary social, economic, and environmental challenges (Mann et al. 2016; Jain et al. 2020).

2.1.3 Nuanced data challenges faced in low and middle-income contexts

Whilst high quality datasets are a significant gap in the ML landscape, in Africa, for example, there are deeper obstacles for researchers who are engaged in AI research and development. These challenges include infrastructure limitations, legal complexities, access to technology, data ownership ambiguities, know how, and access to software.

Researchers in LMICs need to contend with inadequate infrastructure for creating datasets and developing models (e.g. insufficient computing power), and a lack of relevant skills. This hinders their ability to compete with major tech corporations (Rubinfeld and Gal, 2017). Complicating matters further, legal barriers like data protection and privacy laws pose obstacles to researchers seeking access to relevant data for their AI projects. Additionally, the intricate landscape of data ownership, especially in cases involving multiple stakeholders or non-nationals, presents challenges that necessitate resolution. Addressing these legal uncertainties is pivotal to facilitating unhindered data access for research purposes (Rubinfeld and Gal, 2017).
Furthermore, a lack of interoperability between datasets from disparate sources poses challenges to synthesizing and analyzing data (Mupangwa et al., 2020). Similarly, the scarcity of accessible, high-quality analytical tools and algorithms hampers the progress of AI research (Mutai et al., 2021).

Big Data hardware capacity is particularly concentrated within developed nations, and while this digital divide has always been in existence, it is more pronounced in the AI space and has further contributed to the issues faced by researchers in LMICs (Carter et. Al, 2020). Equitable access to telecommunication infrastructure and cloud resources is pivotal to distributing information capacity more uniformly across societies and this needs to be addressed by establishment of social ownership over telecommunication access (Hilbert, 2016). While the digital divide presents a challenge, the mere provision of resources is insufficient; researchers require comprehensive training, support as well as access to various software to effectively utilize these resources (Bezuidenhout et al., 2017). Furthermore, disparities exist in the awareness and adoption of Free and Open Source Software (FOSS) alternatives between low-income and high-income countries (Silva et al., 2023). Despite an interest in embracing and developing FOSS solutions, harnessing their potential requires heightened awareness, knowledge dissemination, comprehensive training, and robust support systems (Vermeir et al., 2018).

2.1.4 Predicted trends in AI over the next decade(s) and anticipated scale of data demand

The number of open data initiatives in developing countries remains limited. It is predicted that the coming years will probably see a large increase of open data initiatives in LMICs. Both civil society organizations and external partners of developing country governments are encouraging the use of open data to increase transparency, accountability and citizen participation (Solís and Zeballos (ed.) 2023; van Belle et al. 2018). In particular the Open Government Partnership is promoting open data initiatives in developing countries (van Belle et al. 2018; Schwegmann 2012; Solís and Zeballos (ed.) 2023).

Predicting the exact trajectory of AI over the next decade is challenging given its dynamic nature, but several trends are anticipated. The coming years are anticipated to witness a substantial rise in open data initiatives within developing countries, as civil society organizations and external partners of these countries advocate for the utilization of open data to enhance transparency, accountability, and citizen engagement (Garcia, 2013).

At the advent of the rapid scale of generative AI there is growing need for investment in the development of unbiased datasets, particularly in the Global South. There are few funders with a specific focus on supporting the development, expansion and maintaining of these datasets. Lacuna Fund offers a unique contribution in this area.

2.2 About Lacuna Fund

In 2020, Lacuna Fund was borne from a funder collaboration between The Rockefeller Foundation, Google.org, Canada's IDRC and GIZ, to support the development of unbiased datasets. Lacuna Fund identified multiple challenges that exist in low and middle-income contexts preventing ML:

- There is a lack of datasets for training and evaluation relevant to low and middle-income contexts and under-represented communities in high-income countries.
- Existing datasets are often outdated, biased, missing key information, or not representative of the country's population or context.
- Limited regulatory frameworks exist to guide new dataset creation for ML and AI
- An inequitable distribution of capital for dataset creation artificially limits creation of open-source datasets relevant to marginalized underserved contexts.

Lacuna Fund provides data scientists, researchers, and social entrepreneurs in low- and middle-income contexts globally with the resources required to create training and evaluation datasets that address urgent problems in their communities. Lacuna Fund's mission is to fund the creation, expansion, and maintenance of equitably labeled datasets that enable the robust application of ML tools of high social value in the Global South and for underserved populations globally. The purpose of Lacuna Fund is to enable researchers and data scientists to create, expand, and maintain ML datasets that are accessible to- and used by developers to achieve benefits relevant to the context, culture, crops, and languages of communities currently under-represented in global data. The Fund's specific goals are to:

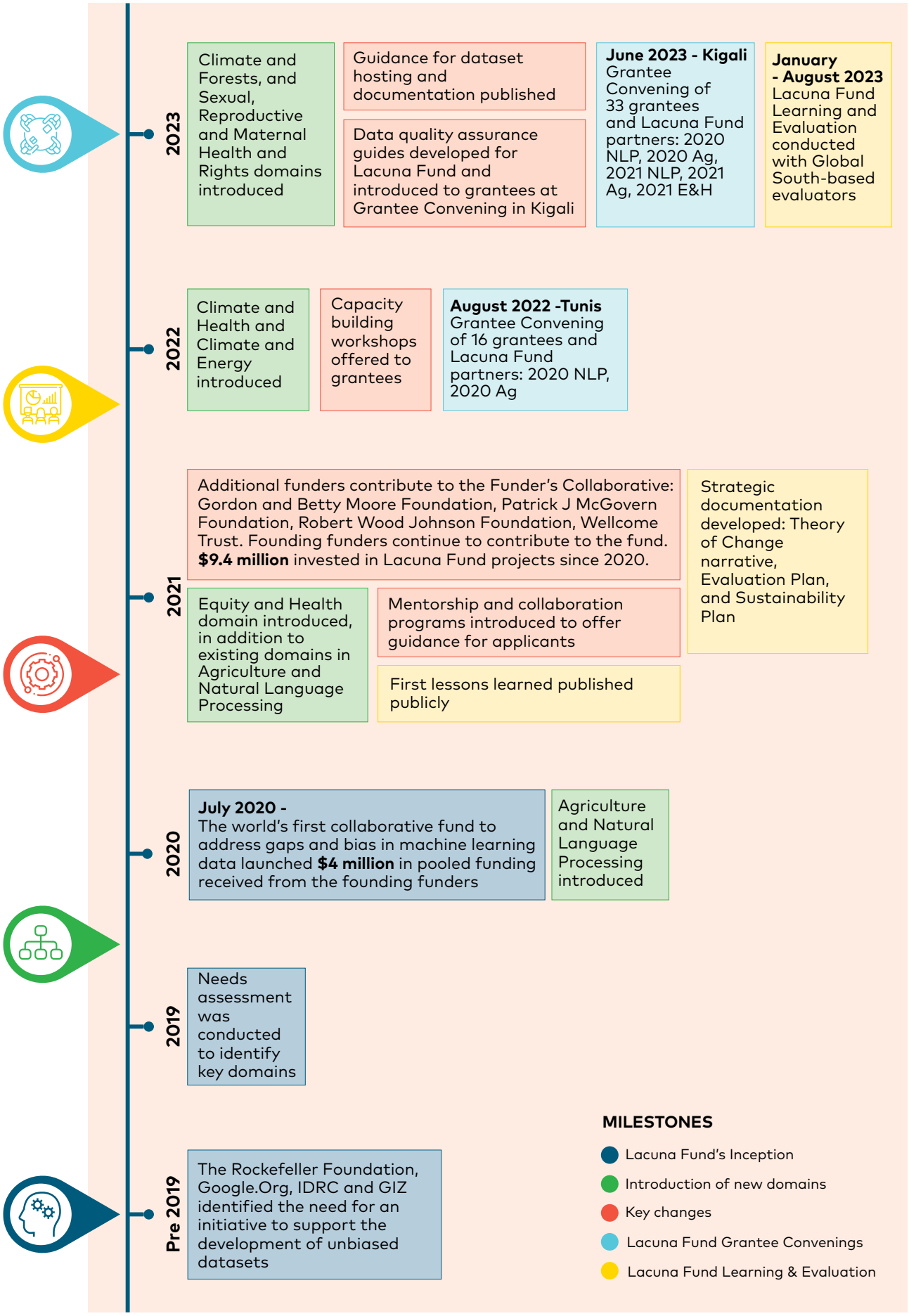
Contribute an investment of funds to institutions to create, expand, and/or maintain datasets that fill gaps and reduce bias in training data used for ML

Enable underserved populations to take advantage of advances offered by AI

Deepen understanding of how to most effectively and efficiently fund development and maintenance of equitably labeled datasets

Lacuna Fund developed principles to promote contextually appropriate, less-biased, more representative ML datasets. These principles are **accessibility, equity, ethics, participatory approach, quality, and transformational impact**. Lacuna Fund seeks to apply these principles in each of its governance structures in the funding cycle. Key milestones in Lacuna Fund's journey are shown in Figure 1.

Figure 1. Timeline of Lacuna Fund inception and key points 2020-2023



2.2.1 Lacuna Fund Governance

Guided by ML professionals worldwide, Lacuna Fund is designed for and by the communities it will serve. Lacuna Fund's governance is a multi-tiered structure.

Lacuna Fund is governed by a representative Steering Committee comprised of 5-9 members. The committee provides strategic direction and oversight for the Fund, working to ensure the Fund's focus, impact, and growth.

Domain-specific Technical Advisory Panel (TAP) members provide technical guidance for the Fund, advising on the focus of requests for proposals within each domain area, selecting proposals, and distilling learnings from the funding process.

The Secretariat provides backbone support to the Fund as a whole, including managing the RFP process and distribution of Funds. Meridian Institute serves as Secretariat for Lacuna Fund.

Lacuna Fund receives funding from a pool of funders. The Funders Collaborative, comprised of current contributors to Lacuna Fund, shares expertise and experience in funding machine learning and AI initiatives and provides input on strategy and priorities through their representatives on the Steering Committee.

2.2.2 Theory of Change

A high-level theory of change exists for Lacuna Fund that is rooted in its fundamental purpose to fill gaps in training and evaluation datasets for machine learning so that AI models can be equitable. According to the theory of change, Lacuna Fund aims to address a lack of datasets for training and evaluation of ML models relevant to low and middle income countries and underserved communities in high-income countries; outdated and unhelpfully biased datasets with missing information and not representative of country population or context; limited regulatory frameworks to guide new dataset creation for ML and AI; as well as an inequitable distribution of capital for dataset creation that artificially limits creation of open-source datasets relevant to underserved contexts.

The ultimate impact of the fund is to create **AI-powered data processing that is less biased and benefits that are more equitably distributed** (source: Lacuna Fund Theory of Change, Annex A.1).

2.3 Evaluation purpose

In 2021, the Secretariat of Lacuna Fund contracted a Monitoring and Evaluation (M&E) consultant to develop the mentioned theory of change, evaluation plan, results framework and to draft data collection tools. Furthermore, the Secretariat was recommended to conduct an evaluation and implement M&E practices to support learning and improvement. In 2023, Data Innovators was contracted to conduct an evaluation of the Fund (2020-2022) and support learning efforts.

The purpose of this evaluation was to assess what aspects of Lacuna Fund have effectively and efficiently enabled the creation, expansion, and maintenance of representative and unbiased training datasets for ML; to examine challenges experienced by stakeholders in Lacuna Fund processes; and to provide recommendations for the improvement of Lacuna Fund.

The evaluation was guided by Lacuna Fund's evaluation plan and Lacuna Fund's recommended utilization-focused approach.

The evaluation aimed to foster reflection and continuous learning in the following focus areas:

- Internal processes relating to the "how" of Lacuna Fund's work, including its governance structures, and operating and funding processes.
- Outcomes - the intended and unintended results of Lacuna Fund's activities as it pertains to its grant-making domains (NLP, Agriculture, Equity and Health, Climate and Health, Climate and Energy).
- Influence - the ripple effects of Lacuna Fund's strategies and the individuals and institutions who are influenced by it.

The following additional focus areas were subsequently added as part of the evaluation's focus areas:

- Principles - the extent to which Lacuna Fund is guided by its principles.
- Landscape - to understand what is changing in the context of Lacuna Fund's work and what strategies are helping grantees to sustain their datasets and keep them evergreen.

Section 2.3.1 below provides an overview of the evaluation questions relating to the previously described key areas of the evaluation.

The evaluation was focused on addressing the evaluation questions.

This report documents the evaluation purpose and focus, methodology, key lessons and insights, and recommendation for the improvement of Lacuna Fund processes, outcomes, and influence. The document is prepared for use by internal stakeholders and funders. A summary infographic has been developed for dissemination.

2.3.1 Areas of Inquiry and Evaluation Questions

PROCESS

1. How and to what extent are the Fund's principles realized in its decision-making structures and funding processes? Do grant decisions reflect these principles?
2. How do Lacuna Fund's governance and processes contribute to improved access to funding for work that is underrepresented and under-resourced, including for researchers based in low and middle-income contexts?
3. What changes can Lacuna Fund make to its processes to continuously improve its ability to achieve its desired outcomes and influence?

OUTCOMES

4. How and to what extent do complete datasets align with Lacuna Fund principles? To what extent do completed datasets address the equity or representation gaps they initially targeted?
5. How and to what extent are Lacuna Fund-supported datasets being used, maintained, and updated to stay accurate and current? Who is using the datasets?
6. Who is likely to benefit from these applications? To be left out of benefits or even harmed by these applications?

INFLUENCE

7. How and to what extent does Lacuna Fund's work influence other funders of datasets? Are these investors making the work they fund public and are they considering dataset bias?
8. How and to what extent do indirect benefits or unanticipated challenges accrue to the researchers funded by Lacuna Fund or to their research institutions because of their involvement in the projects?

LANDSCAPE

9. What is changing in the context of Lacuna Fund's work?
10. What strategies are helping grantees keep their datasets evergreen or sustain their usability in these everchanging contexts?

3. Methodology

3.1 Approaches and methods

Lacuna Fund's evaluation was conducted between January to August 2023. It was guided by ten main evaluation questions (Section 2.3.1), with an additional eight secondary questions. Mixed methods and a utilization-focused evaluation approach was used. Whilst four methods were in the initial protocol of the evaluation, an opportunity to gain more insights through the Grantee Convening session emerged and was included in the evaluation findings.

Data collection instruments were designed for each method; the five data collection methods used were as follows:

3.1.1 Desktop review

The desktop review process was conducted in two stages, (1) a review of the existing Lacuna Fund's evaluation plan, outcomes, indicators, data collection methods and instruments, as well as the theory of change; (2) an ongoing review of internal documents that influence decision-making processes (e.g surveys, evaluation sheets, reports, dataset websites etc.). In conjunction with this process, a literature review was also conducted. More than 130 documents, databases, websites and literature were reviewed.

3.1.2 Stakeholder engagements through sense-making workshops and staff reflections

A total of three online participatory sense-making workshops were conducted with 11 Lacuna Fund representatives (Secretariat (4), the Steering Committee (4) and selected Technical Advisory Panel (TAP) (3) representatives). These workshops were conducted to prompt reflections and gain a deeper understanding of Lacuna Fund processes, lessons, and cases of influences and sustainability; as well as to sense-check the preliminary insights from other data sources. A staff reflection session was conducted with a Secretariat staff member to source reflections on their day-to-day operations of the Fund that provided key insights into the collected data.

3.1.3 Interviews: grantees and bellwethers

Semi-structured interviews were conducted with two sets of stakeholders: grantees, and bellwethers. Twenty of the sample of 30 grantees were interviewed, from Agriculture (4), Climate and Health (1), Climate and Energy (1), Equity and Health (4) as well as the NLP (10) domains. They provided an in-depth understanding of their experiences with Lacuna Fund processes, lessons learned, the benefits they encountered as well as the contexts in which they implemented their research projects and are based.

Thirteen of the sample of 15 bellwethers were interviewed. The bellwethers were identified thought leaders who provided high-level perceptions of the role of Lacuna Fund and/or similar initiatives of ensuring accessible datasets, and views on the landscape and shifts in equitable data access. The respondents were from the domains: Climate and Energy (2), Agriculture (1), Health and Racial Equity (2), NLP (2), Energy (1), Information Technology (1) and four representing cross-domain views. Of these, one was a Steering Committee member, three were TAP members, one was both a TAP member and grantee, two provided advisory support and four provided contractual support to the Fund.

3.1.4 Surveys: dataset users and grantees

A survey was disseminated to dataset users, and another one to grantees. These surveys were shared through Lacuna Fund's mailing contact list and social media platforms (LinkedIn and Twitter). The dataset user survey was disseminated in **four languages** (English, French, Spanish and Portuguese). 62 responses were targeted for the user survey; 48 responses were completed, giving a satisfactory 77% response rate.

For the grantee survey, grantees were included that were not reached for interviews as well as applicants who did not receive grants due to limited resources but whose proposals received high marks. Twenty-one of a sample of 74 grantees completed the survey by the end of the Grantee Convening at the AfricaAI Conference, a 28% response rate. In conjunction with the grantee interviews, this response rate was not necessarily a limitation.

3.1.5 Grantee Convening session

A 90-minute participatory action-mapping session was conducted with 17 grantees at the Grantee Convening, on June 16, 2023. This was done to elicit additional insights from grantees that contribute to the questions around processes and outcomes of Lacuna Fund's work, as well as to provide recommendations they have towards more equitable processes and transformative impact, focusing on four themes: (1) Lacuna Fund's application and grant disbursement process, (2) the sustainability of datasets, (3) the use of datasets and populations who use them, and (4) researchers' benefit from involvement in funded projects.

3.2 Analysis

Four analytical approaches were used for the data analysis of this evaluation: sense-making, descriptive analysis, deductive analysis and thematic analysis. Upon the completion of the data collection phase, the research team collated data from the various datasets and cleaned the data in preparation for analysis. Once this was completed, the interviews, surveys and grantee database data were coded for analysis, with an ongoing quality assessment to ensure reliable data. Trends and themes were triangulated and analyzed from different sources.

3.3 Learning and feedback

The preliminary findings generated from the various data collection points were documented in two learning briefs that highlighted emerging insights and recommendations. The Secretariat has shown immediate responses to some of the recommendations from the brief such as (1) discussion of reframing of the principle of Transformational Impact and measurement thereof, (2) the release of a resource guide to support grantees with technical knowledge on publishing datasets, (3) creating a platform for inter-project collaboration amongst the Grantees by encouraging them to share their contact details for future purposes and (4) ensuring there is a participatory approach in the fund's processes by allowing grantees to contribute to additional recommendations through the Grantee Convening session.

3.4 Limitations

The data collection was mostly conducted online, and whilst this may be convenient, it may have caused limitations that excluded some participants. When the interviews were conducted, the different time zones were considered because the grantees and bellwether interviewees were based at various geographical locations. Numerous times the network connectivity was also a challenge for interviewees. The interviews were mostly conducted in English, with one French interview. The language barrier may have limited participants from participating in the interviews. Geographical information was not collected on the dataset surveys, and therefore were unable to confirm if the respondents were based in low and middle-income contexts.

3.5 Ethical considerations

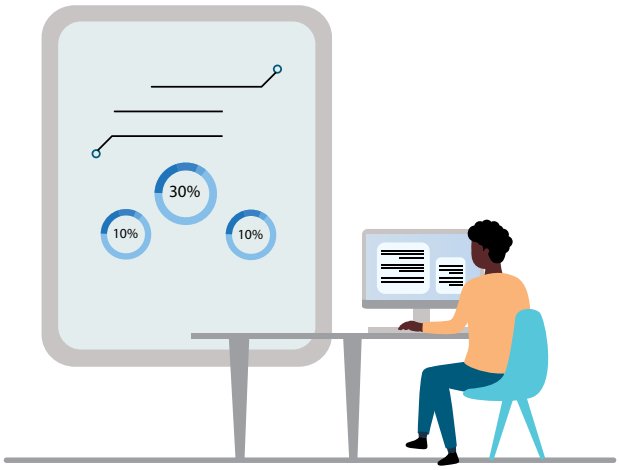
The preliminary findings generated from the evaluation adhered to some ethical considerations including the following: Informed consent was sought from all respondents participating in interviews and surveys; this was done either through signed written consent forms or agreed verbal consent before recording interviews. This process was voluntary, and respondents were made aware of their right to opt out of an interview, at any point. Respondents were made aware of our voluntary process of collecting confidential information such as gender and ethnicity; and highlighted that no identifying details would be included in the final report. Respondents were made aware of their right to access the results of the study or have an abstract made available on request. In the cases where consent was not provided by the potential participants, the evaluation team did not proceed with the data collection efforts. One grantee interviewee opted out of being recorded, but agreed to continue with the interview.

4. Findings

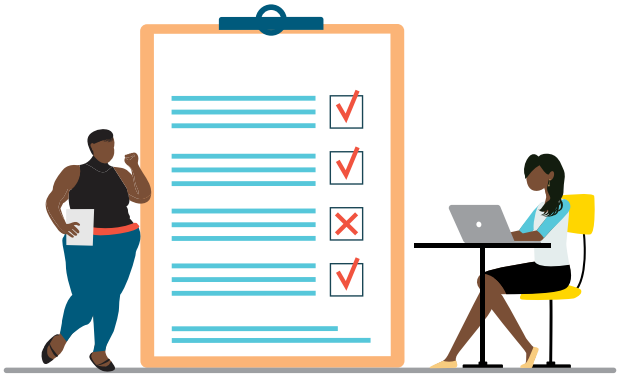
The findings include a consolidation of insights in response to the evaluation questions. Each section incorporates data triangulated from the described methods. These findings form the basis for recommendations and suggested updates to the results framework.

The section is structured according to the following sub-sections:

Overview of emerging outcomes



Evaluation areas of inquiry



Lessons learned

Below is a list of the primary findings and impact statements that reflect what Lacuna Fund has achieved over 2020-2022.

4.1 Overview of Emerging Outcomes

Since its inception, Lacuna Fund has achieved a broad array of impact in the field of dataset development in low and middle-income contexts, and made continuous improvements to its processes as a grant-maker for public good. The most notable achievements since 2020 are detailed in this section. These include evidence that Lacuna Fund:

is embedding a principles-based approach to collaborative funding that can be scaled and replicated in other funds

has increased access to funding for datasets in low and middle-income contexts

is increasing equitable data usage and supporting underserved researchers and data scientists

is creating connections amongst grantees and networks for knowledge sharing and further impact and reach

is prioritizing high-impact domains that will affect development goals

is elevating the reputation of pioneers and change-makers in the landscape of dataset development for ML

is establishing a trusted brand as a funder and source of quality datasets and research

Embedding a principles-based approach to collaborative funding

All principles are **intentionally embedded across Lacuna Fund's processes and reach**. Investment in a principles-based approach is intended to ensure consistency across Lacuna Fund's implementation over time and instill the principles within the AI ecosystem and funders of datasets - influencing shifting norms in the field. The principles are visible through the Steering Committee and TAP Member compositions, guiding documents and webinars offered during RFPs, and project profiles of grant recipients.

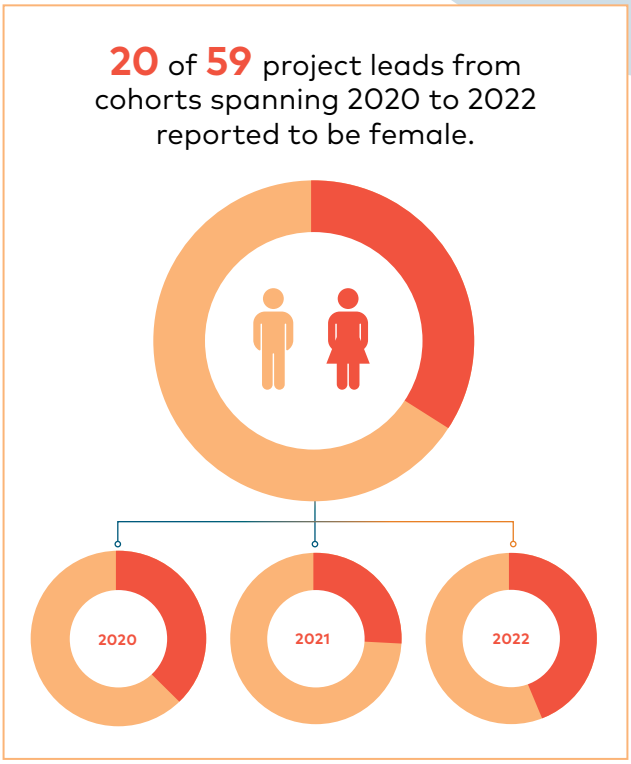
All funded projects are ranked high in the incorporation of the principles at the proposal stage. However, equity and impact must be deepened in the proposals selected for funding. For example, the ratio of females to males remains uneven with 20 of 59 project leads from cohorts spanning 2020 to 2022 reported to be female (Figure 2).

In contrast, 60% of project leads are young professionals in the early phases of their careers. Lacuna Fund has made notable efforts to improve measurement of equity based on inclusive and non-restrictive gender and other demographic categories.

Equity is also addressed in the projects funded particularly by enhancing representation and inclusion through creating new datasets or expanding existing ones. Nineteen grantees shared that their project contributed to improved equity or better representation for underserved populations.

All projects are being run in low and middle-income contexts. 41% (24 out of 59) of lead grantees are not located in the research

Figure 2. Gender representation of project leads



60% of project leads are young professionals in the early phases of their careers.

100% of projects are located in low and middle-income contexts.

country, and all 24 partnered with in-country researchers. On the other hand, 35 of the lead grantees are based in the country where the research is located, 12 of these did not list any research partners at the proposal stage, and the rest have partnered with researchers spanning multiple geographic regions. Lacuna Fund has acknowledged the need to focus on context and not country only, ensuring that projects are intended to broaden reach and impact in underserved communities, areas and other groups that are not geographically confined.

To embody the principle of a **participatory approach**, the Steering Committee, TAP member and Secretariat composition have been deliberately inclusive of diverse demographics and field experts from respective domains. The Steering Committee meets twice a year and makes decisions by consensus, and the Secretariat meets bi-weekly while frequently responding to ad hoc support requests and adapting or developing further systems and tools where needed. Further to this, domains funded are informed by the grantees themselves through assessments and evaluations of what is emerging as key in the landscape of ML and datasets.

Increased access to funding for datasets in low and middle-income contexts

Lacuna Fund is one of few grant makers funding the development of equitable datasets. Three other funders were identified to have a specific fund or focus on reducing bias in datasets, and about 60 funders invest in areas of model development, research, ethical AI and events. Lacuna Fund is made up of a collective of

Figure 3. Funding key data points

$9.4 million collectively invested in Lacuna Fund projects between 2020-2022
The fund has continued to disburse funds in 2023

59 projects
funded to serve the domains of NLP, Agriculture, Equity and Health, Climate and Health and Climate and Energy

100%
projects funded benefitted low and middle-income contexts in 54 countries globally

funders that doubled from four in 2020 to eight in 2021 with the inception of the Equity & Health domain RFP.

The growth in funders is indicative of emerging outcomes of new investment entering the field. Funding was distributed based on findings from a needs assessment to determine the relevant domains in 2019 and in alignment with areas of funder focus.

In 2021, two Equity and Health domain projects based in USA were funded. Both benefited low and middle-income contexts in the USA. The respective project titles were "Machine Learning from Real Patient Outcomes to Reduce Racial Disparities in Chronic Pain" and "ACCELERATOR - Fair Clinical Machine Learning Data Training Consortium". The first project addresses low and middle-income contexts by recommending cost-effective treatments, tailoring personalized plans considering financial limitations, and providing culturally sensitive guidance, thus promoting equitable access and

outcomes. Collaboration with community organizations and advocating for policy changes further enhance its impact in these contexts. The second project seeks to advance fair ML in healthcare. By fostering collaboration among stakeholders, it aims to ensure equitable representation in clinical data, develop unbiased algorithms, and promote responsible AI deployment in medical contexts, ultimately benefiting underserved populations.

Multiple developing countries with rapidly growing digital economies in both public and private sector hold recipients of Lacuna Fund grants. Kenya and Nigeria each hosted six lead grantees and a greater scale of projects compared to other countries represented (Kenya 10 projects, Nigeria 11 projects).

Lacuna Fund has a **diverse spread of funding across organization types** including academic and research institutions (32% of funds), non-profit organizations (37%), researchers and professionals in public sector entities and private for-profit companies (31%). **Expansion of the pool of grassroot researchers** receiving funds may be explored to extend reach to underserved data scientists and researchers.

Lacuna has pivoted in some areas to address barriers to inclusivity of the fund, such as the adapting the focus from "labeled" datasets to "training and evaluation datasets" or "datasets for training and evaluation of machine learning models". The broader scope allowed for more variation in project types and grantees that are still aligned with the purpose of the fund.

Overall, the increase of funding to low and middle-income contexts has enabled the much-needed broadening of representative research in ML and AI, through locally-based researchers and data scientists.

Increasing equitable data usage and expanding growing underserved researchers and data scientists

The datasets alone will not create systemic change or shifts in norms around equitable data. The usage of the datasets and projects that spawn from the initial grant are signals of positive outcomes of Lacuna Fund. There is evidence to indicate **that usage and engagement with the datasets are emerging.** The results thereof are yet to emanate from these gains. Investigating the cases and stories of the use and benefits will be useful to understand the extent of these outcomes.

Six grantees confirmed knowledge of their datasets being used by other stakeholders. All 17 completed projects host over 30 datasets that are published on popular open-source platforms such as ML Hub, Zenodo, Inalco Bomba Reference Corpus, GitHub and Harvard Dataverse. **The requirement to publish on open-source platforms is a step towards encouraging other researchers to engage with the data, and grantees to expand on their projects.**

A dataset containing data on malnutrition of children in Chile has been made locally available, but not publicly available, because it contains identifying information that is critical to its continued use by other government departments in Chile, for the benefit of the respective children in the dataset.

More than 70% of grantees are working on building new datasets, and some grantees are expanding existing datasets to make them more representative and inclusive. The funding has also further allowed grantees to acquire

Figure 4. Data usage and access data points

6 grantees cited various under-represented communities that are benefitting from their work, including dataset users, farmers, students and researchers

17 projects

>30 datasets completed and published on Harvard Dataverse, ML Hub, Zenodo, Inalco Bomba Reference Corpus and GitHub

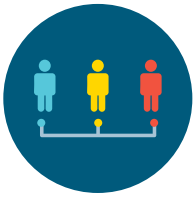
published data sets downloaded **>407k times**

cited **19 times**

Kencorpus dataset on **3 Kenyan Languages** i.e. Kiswahili, Dholuo and 3 Luhya dialects (Lumarachi, Logooli and Lubukusu) have been downloaded **>312k times** (as of August 2023)

equipment, improve their services, and partner with stakeholders which would influence furthering these data scientist and researchers' work in equitable data.

Beyond the direct project value, grantees report **personal growth, skills developed and visibility** as additional benefits of the fund. **Two grantees reported obtaining additional funding and new projects** as a result of their Lacuna Fund-funded projects.



Connecting grantees to one another, and to key networks for knowledge sharing and further impact and reach

Grantees are inspired to address challenges in their local contexts, and they recognize that data is a powerful tool in that regard. They are also inspired by one another's work and have suggested that a platform is established that allows them to collaborate and share knowledge, in addition to the annual convening. They want to stay connected and be able to ask questions more frequently and are willing and eager to coordinate this for themselves. Lacuna Fund responded to this emerging insight during the evaluation by establishing a Slack channel. This opportunity for **interconnection and learning loops amongst grantees is a multiplier factor** as it may strengthen the quality of data, sustainability of projects and create new opportunities.

Eight of 20 grantees reported improved networks as a positive outcome of their projects. Three grantees shared that new partnerships were formed with private sector companies and government (e.g. Orange, Google, Buganda Kingdom).

Lacuna Fund grantees are attending and participating in key and influential conferences where they are able to showcase their work, engage with potential users of the datasets and connect with others in the field. Examples of conferences attended are the Deep Learning Indaba 2022, Data Science Africa 2022 and the AfricaAI Conference 2023.

Sustainability planning is a requirement for all grantees. Thirteen grantees provided details on sustainability in final reports (submissions by March 2023). And many grantees reported that they have developed sustainability plans, though the detail and longevity of the plans was not assessed. The most common strategy described to sustain the usability of datasets being employed by grantees is model building. In addition, partnership was identified as a key solution to sustainable datasets by grantees at the Grantee Convening.

Figure 5: Partnerships formed by grantees

New partnerships formed as a result of their Lacuna Fund projects

- Project commenced with **Orange** on automatic speech recognition and has interest from **Google**.
- "...the partnership itself is a first positive point of this success story, because you have a large multinational company, a small company like ours, and a higher education institution like EPT, so in terms of partnership and innovation in this sector, it's quite unique in Africa, and I think that's a success story."

Figure 6: Conferences attended by grantees

6 grantees in the NLP domain and **2 in the Agriculture domain** engaged in the Deep Learning Indaba 2022 through conference presentations, posters, or demonstrations.

3 grantees from the NLP domain and **1 grantee from the Agriculture domain** participated as official speakers at the Data Science Africa Conference 2022.

Prioritizing high-impact domains

Domains of high impact are sectors or areas where there is a high demand for data, and data usage may play a critical role in addressing development goals such as improving livelihoods through agriculture and reducing the digital divide by integrating local languages in social media and online platforms. The selection of domains is informed by multiple factors - including the initial needs assessment, input from the field (including information from this evaluation), insights from the Steering Committee, and the interests of funders - to ensure relevance to all stakeholders and local needs.

NLP and Agriculture were the two most prominent domains submitting proposals for funding and 55% of funds awarded between 2020-2022.

Eleven of the 20 grantees interviewed are already building models based on collected data and six are integrating the models into real-world applications within the domains of Agriculture (4), Equity and Health (1), Climate and Health (1). Use cases include **crop type mapping, disease classification, outcome prediction,** etc. Since the interviews only included a sample, it is possible that there are more modeling activities underway from other grantees. Nonetheless, these mentions are demonstrable of emerging applications that may have high impact in communities.

Grantees and bellwethers reiterated the importance of continuing to fund NLP and Agriculture domains. They recommend new domains in Disaster, Financial Services, and Transport and Logistics. Climate was identified as a priority domain of interest for researchers and data scientists in 2021 and 2022. As the impact of climate change becomes more blatant, this domain can scale much needed new datasets to accelerate model development for rapid decision-making.

Figure 7: Examples of use cases in Agriculture

Two cases of Lacuna Fund projects addressing challenges in Agriculture (Detail in AI Media, 2021):

Helmets labeling crops
"Rapid point data collection with cameras mounted on the hoods of vehicles combined with crowdsourcing...to improve agriculture monitoring"

Eyes on the ground: providing quality model training data through smartphones
"use smartphones to create a unique dataset of geo-referenced crop images along with labels on input use, crop management, phenology, crop damage, and yields."

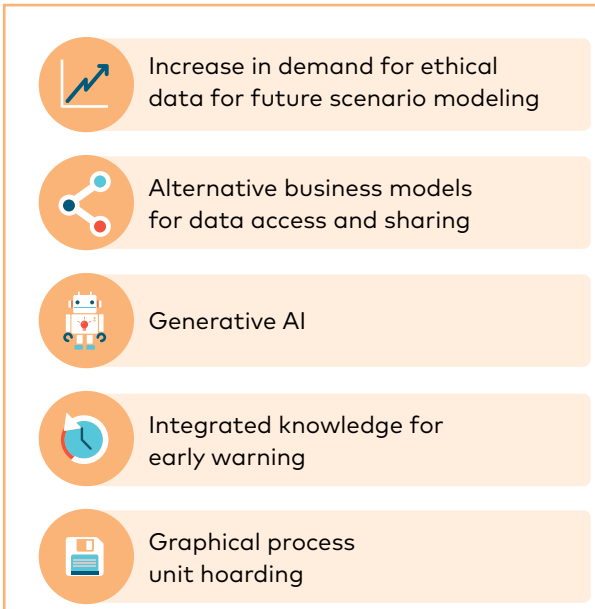
Elevating the reputation of pioneers and change-makers in the landscape of dataset development for ML

Lacuna Fund is increasingly establishing itself, TAP members and grantees as expert resources and partners as evidenced by grantees and bellwethers (including TAP members) and their presence in the main ML communities in LMICs such as Deep Learning Indaba, Data Science Africa and Masakhane. By choosing to fund only projects in low and middle-income contexts, funding projects that have development objectives, the decentralized governance structure and application of lessons learned, Lacuna Fund can **make shifts in "power" in the field of AI.**

> *"It is not uncommon now for AI experts to ask whether an AI is 'fair' and 'for good'. But 'fair' and 'good' are infinitely spacious words that any AI system can be squeezed into. The question to pose is a deeper one: how is AI shifting power?"*
>
> - Pratyusha Kalluri

Lacuna Fund's influence on other funders of datasets is growing amongst key players in the field, and philanthropic funds (eight funders since 2020). There is opportunity to increase awareness of the fund amongst donors investing in other points in the data value or relevant areas (e.g. model development, responsible AI). Grantees and bellwethers perceive the level of influence on funders as low at this stage. They also feel that Lacuna Fund could partner with more African philanthropic organizations. Some grantees have shared that they found alternative funding for modelling subsequent to the dataset development. There are few other funders that have a particular focus on unbiased datasets, or those that exist are not easily found.

Figure 8. Five trends in the context of Lacuna Fund's work include



Field experts shared five trends in the field of AI and ML which may influence Lacuna Fund and stakeholders in the future. These trends (listed in Figure 8) are areas that Lacuna Fund must be aware of and consider the implications for strategy and sustainability. As the trends unfold, so will the need to reassess Lacuna Fund's influence and role. Lastly, the cross-sectoral lessons from implementing the fund and funding model in itself are pockets of knowledge that Secretariat can use to influence other funders and elevate the reputation of the fund.

Establishing a trusted brand and source of quality datasets and research

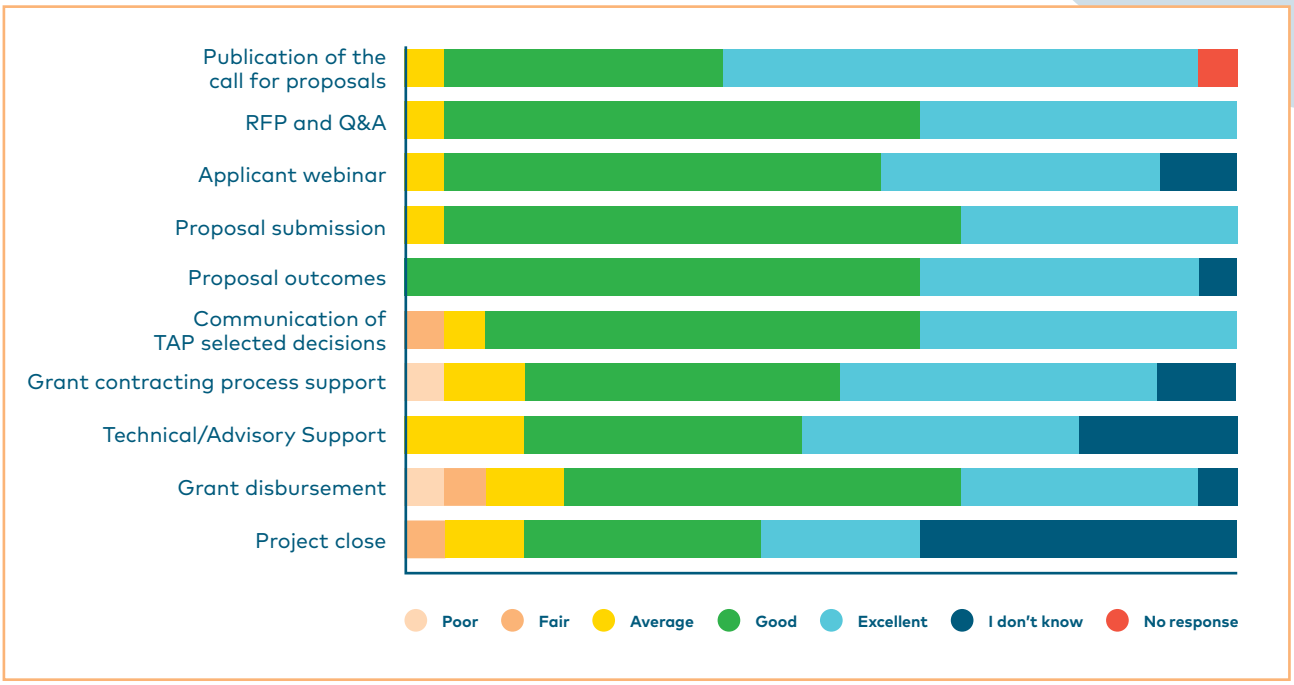
Feedback on grantee and data user experience is important to improve outreach and funding practice. It is also an indicator of trust in the brand and positive association with Lacuna Fund. There is a **high amount of positive feedback** coming from grantees and users of Lacuna datasets who reported that Lacuna Fund's platform met their expectations and they would recommend it to colleagues. The users' net promoter score of nine out of 12 indicates that a higher proportion of users surveyed are likely to promote or recommend the funded datasets. Grantees were unanimous in saying that they would recommend Lacuna to others, and 11 out of the 20 interviewed already have.

Ease of communication and processes, and the professional, accessible and effective support received, were cited as reasons to recommend Lacuna to other researchers. Further to this were the opportunities that locally produced data present for research, to both kick-start and further locally relevant research topics. Grantees reported that the fund Secretariat had **efficient turnaround times, delivered according to their expectations, and that the processes are fair** (Figure 9). Some discrepancies were observed between first and second round of funding release, where the first was found to be efficient but the second had time delays and lacked support for a few grantees, resulting in the observed dissatisfaction. From communication of TAP decisions onwards, the remaining processes were identified as areas of improvement.

> *"They [LF] actually came back when I proposed my grant and went beyond the call. They came back and asked if I wanted more funding, which was unique. So that was very good experience. Once the grant was awarded, my relationship or my interaction with them became very limited, and it was even harder to understand what they want us to do in terms of the reporting...The reporting is different from what I'm familiar with. Just being a challenge on making sure that we're doing the reporting. The money came quite late. There's some amount of communication gap where they send something and then just go cold and you don't hear back from them."*
> - Grantee 1

Notwithstanding the highlights in the emerging outcomes above, as a pioneer and currently sole grant-maker of dataset development in the landscape of ML, areas of further learning and improvement were also identified for Lacuna Fund's consideration. The next section expounds on the above highlights and relays suggested areas to focus on for further growth and an increase in impact with specific regard to the evaluation questions explored.

Figure 9: Grantee satisfaction ratings of Lacuna Fund processes



4.2 Evaluation areas of inquiry

Evaluation questions were posed to delve into understanding achievement and lessons in four areas: process, outcomes, influence and landscape. In this section, a brief on the status of Lacuna Fund in response to the question is provided, in addition to sharing record of key data points. Though outcomes are detailed in the previous sections, this section summarizes additional data points and findings.

PROCESS

1. How and to what extent are the fund's principles realized in its decision-making structures and funding processes? Do grant decisions reflect these principles?

In summary
With the governance structures and shifts to accommodate more accessibility, along with careful selection of principles-aligned proposals by TAP members and constant adaptive support offered to grantees by the Secretariat, the fund's principles are clearly realized across processes and reach.

On a scale of one to three where one dot denotes "average", two dots denote "above average", and three dots denote "good", the ranking below demonstrates the extent to which the principles are realized and examples thereof are explained in the section that follows.

The principle of impact is an exception as datasets have not yet matured enough to reflect this principle. The principle of quality is yet to be assessed on published datasets by data scientists who can evaluate if the required standards are being met, however, the TAP does evaluate the incorporation of these aspects in proposals received.

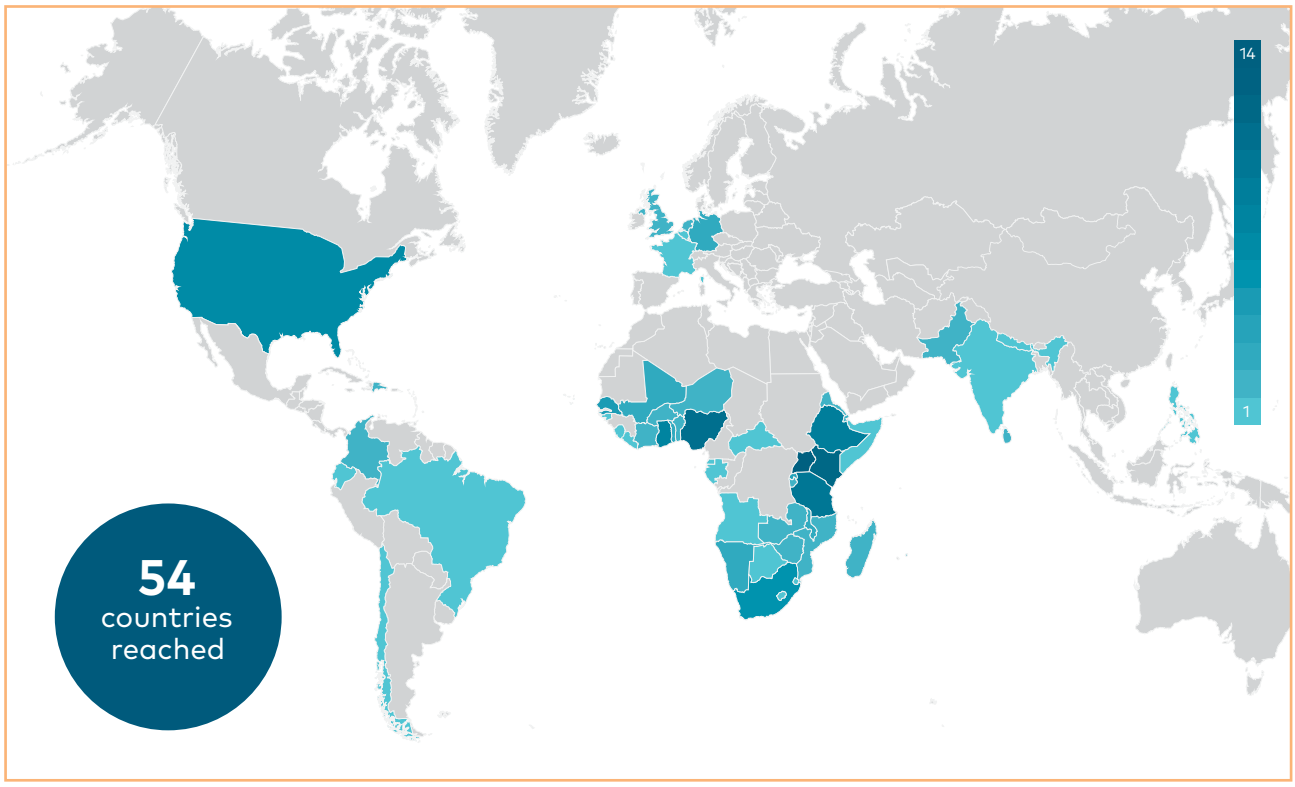
KEY DATA POINTS

ACCESSIBILITY	EQUITY	ETHICS	PARTICIPATORY	QUALITY	IMPACT
Dispersing calls for proposals in multiple languages and requiring grantees to make datasets openly available.	Grantees are filling data gaps and making ML and AI more equitable. Increased funding representation of LMIC regions over time and improved female ratio ofgrantees.	RFPs require applicants to describe steps to ensure ethical approaches to data collection. Lacuna Fund Conflict of Interest policy made available to grantees.	RFPs ask applicants to consult or collaborate with potential data providers and users. Governance structure & processes also participatory and inclusive.	TAPs review applications for quality. A Data Quality Advisor was contracted in 2023 and is reviewing applications, grant reports and datasets, and providing feedback on quality.	It will take time to realize the impact gained by communities. Monitoring of use cases for impact will continue.
● ● ●	● ● ○	● ● ○	● ● ●	● ● ○	● ○ ○

Scope of projects in low and middle-income contexts as a reflection of TAP grant-making decisions

- 100% of all projects are classified as benefiting under-represented or under served populations in 54 countries across the world.
- There are many instances of projects benefitting communities in more than one country. In some of these cases, the project locations will be a mixture of countries with different income levels.
  There are two grants that were awarded to USA-based institutions for work being carried out in the USA through the Equity and Health call in 2021, however, as required by Lacuna Fund, this work benefits under-represented communities in the USA. There are five other projects with multiple data collection locations that include a mixture of countries at higher and lower income levels, including the USA.

Figure 10: Countries where datasets are being collected



Note: Of 7 projects awarded grants in the US, 2 are for the benefit of marginalized communities in the Equity and Health domain, and the other 5 are projects with multiple data collection locations that include countries across different income levels.

Incorporation of Lacuna Fund principles into decision-making structures and processes

- The TAP's technical know-how and overall grant-making process is supported by:
  - a guiding overview document that gives an in-depth outline of their roles and responsibilities, and ways in which the TAP can assist in refining the processes that are already in existence or advise on the creation of new processes that would be valuable.
  - a proposal review guide provides step-by-step instructions and timelines on how members of the TAP should conduct proposal reviews and this process ensures uniformity and reduces bias in how proposals are reviewed.
  - a conflict of interest policy that speaks to further processes that are essential in ensuring that grant provision policies are fair and don't offer any unfair advantages.
- intellectual property policy that provides a broad explanation for how the datasets developed with grant funds will be handled, stored and shared.
- Since inception, RFPs have been made available in 3 languages: English, French and Spanish.
- The Secretariat supports by continuously engaging both applicants and grantees to offer ad-hoc support and clarify any required procedures, as well as meeting bi-weekly to keep abreast of all operations. Major decisions are handled by the Steering Committee bi-annually and their ToRs outline a process for making decisions by consensus.

The principles were also evident during this evaluation process in which immediate responsiveness to emerging insights was often seen, such as:

- a Slack channel was started for grantees to connect with one another and Lacuna Fund's Secretariat more regularly and share learnings.
- an invitation was extended to the evaluation team to attend the Grantee Convening held in Kigali in June 2023 to further elicit recommendations from grantees.

Relevant indicators

Indicators	Target	Actual
1. Perceived degree of alignment between the equity gaps addressed by funded projects (as described by grantee) and the priority representation issues (as established by the TAP)	100%	71%
2. % funded teams based in- country	100%	64% of lead grantees based in-country 100% of projects have in-country partners
3. % funded project designs that include a participatory approach to dataset development	100%	63%
4. % funded project designs that describe how value/ benefits of the project will be shared with local community/data contributors	100%	69%
5. % funded project designs that meet standards on (a) privacy concerns, (b) mitigation of potential for downstream misuse, (c) possible discrimination vectors (e.g., gender), and (d) fair and equitable working conditions, if paid labellers are involved in the project	100%	a) 98%, b) 71%, c) 71%, d) 71%
6. Percentage of funded projects that demonstrate the ability to extend design practices to industry best practice		66%
7. No. of languages RFP is distributed in (and distribution channels)	None yet	3 (English, French, Spanish)
8. Perceptions on quality and use of feedback from grantees about Lacuna Fund		Very good (where 1=poor, 2=average, 3=very good, 4=excellent)
9. Composition of the TAPs reflect industry expertise as well as lived experience in the communities /regions prioritized by Lacuna Fund		100% of TAPs have had representation from both industry and the target region.

[3]The actual values for Indicators 1, 3 & 5 are based on proposal assessments by the evaluators. At a later stage, for a future evaluation, these can be cross-checked against the TAP's scores of the proposals against Lacuna Fund Principles

PROCESS

2. How do Lacuna Fund's governance and processes contribute to improved access to funding for work that is under-represented and under-resourced, including for researchers based in low and middle-income contexts?

In summary

The highlights shown in the emerging outcomes section demonstrate clear and extensive efforts to fund work that is under-represented and under-resourced for researchers based in low and middle-income contexts. With all funded projects benefiting marginalized communities and being led by grantees that are either based in- or are partnered with other researchers in the low and middle-income countries where the research is based. Areas of further improvement in gender representation, grassroots researchers, and increasing accessibility of RFPs and proposal submissions are to be considered by Lacuna Fund.

The principle of impact could not be measured as the datasets have not yet matured enough to reflect impact. The principle of quality is yet to be assessed on published datasets by the Data Quality Advisor who can evaluate if the required standards are being met, however, the TAP does evaluate the incorporation of these aspects in proposals received. With the diversity of applicants, the TAP, Secretariat and Steering Committee of Lacuna Fund play a critical role in achieving Lacuna Fund's mission to benefit communities in low and middle- income contexts that are headquartered within these contexts.

KEY DATA POINTS

Funding allocation to domains in under served and under-represented contexts

- Lacuna Fund has allocated funding to areas that continue to remain relevant and are guided by the needs assessment. Table 3 shows the funding that targeted RFPs have enabled over 2020-2022. Lacuna Fund continues to seek further funding to run calls for proposals in these domains, and others as they emerge.
- Lacuna Fund's governance and processes also contribute to improved access for funding through its global, multi-stakeholder Steering Committee that includes representatives from data science and from the geographies targeted for funding. Lacuna Fund also has multi-disciplinary expert panels (TAPs) with representatives from the regions where funding is granted reviewing and selecting which proposals receive funding.
- As shown in Figure 11, the NLP domain has had the most grants awarded over 2020-2022, and none of them were above US$250,000. In contrast, the more recently introduced domains of Climate and Energy, and Climate and Health have a higher ratio of large funds over US$250,000 being awarded.
- Though the Climate and Health domain received the least sum of funds, it has the highest average project cost based on the four grants awarded to that domain. NLP has the lowest average project cost (Figure 12).
- The communities influenced by this work include marginalized communities, local businesses, policymakers, researchers, government, data scientists and researchers, farmers, health care personnel, native language speakers, small and large-scale farmers, and plant breeders, to mention a few.

Table 3: Funding* awarded to each domain over 2020-2022

	2020	2021	2022	Total
NLP	$1.2M	$1M		**$2.3M**
AG	$1.3M	$900k		**$2.2M**
E&H		$1.6M		**$1.6M**
C&H			$1.2M	**$1.2M**
C&E			$2.1M	**$2.1M**
Total	**$2.5M**	**$3.6M**	**$3.3M**	**$9.4M**

***Note: All funds stated in this report are approximate values.**

Figure 11. Number of grants and grant size per domain

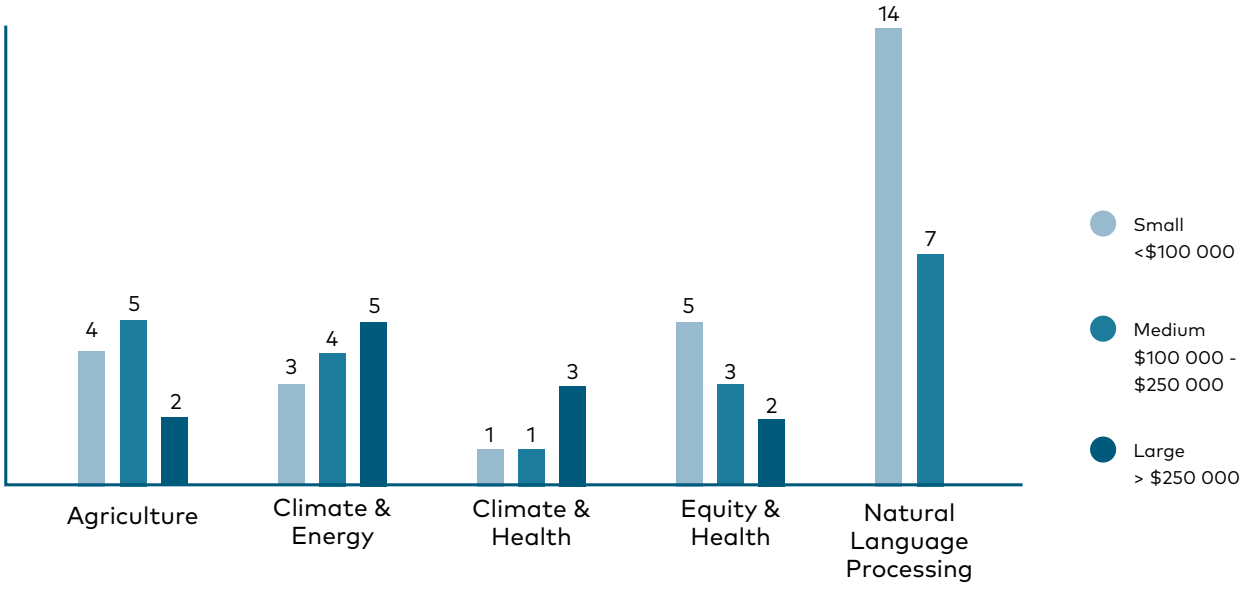
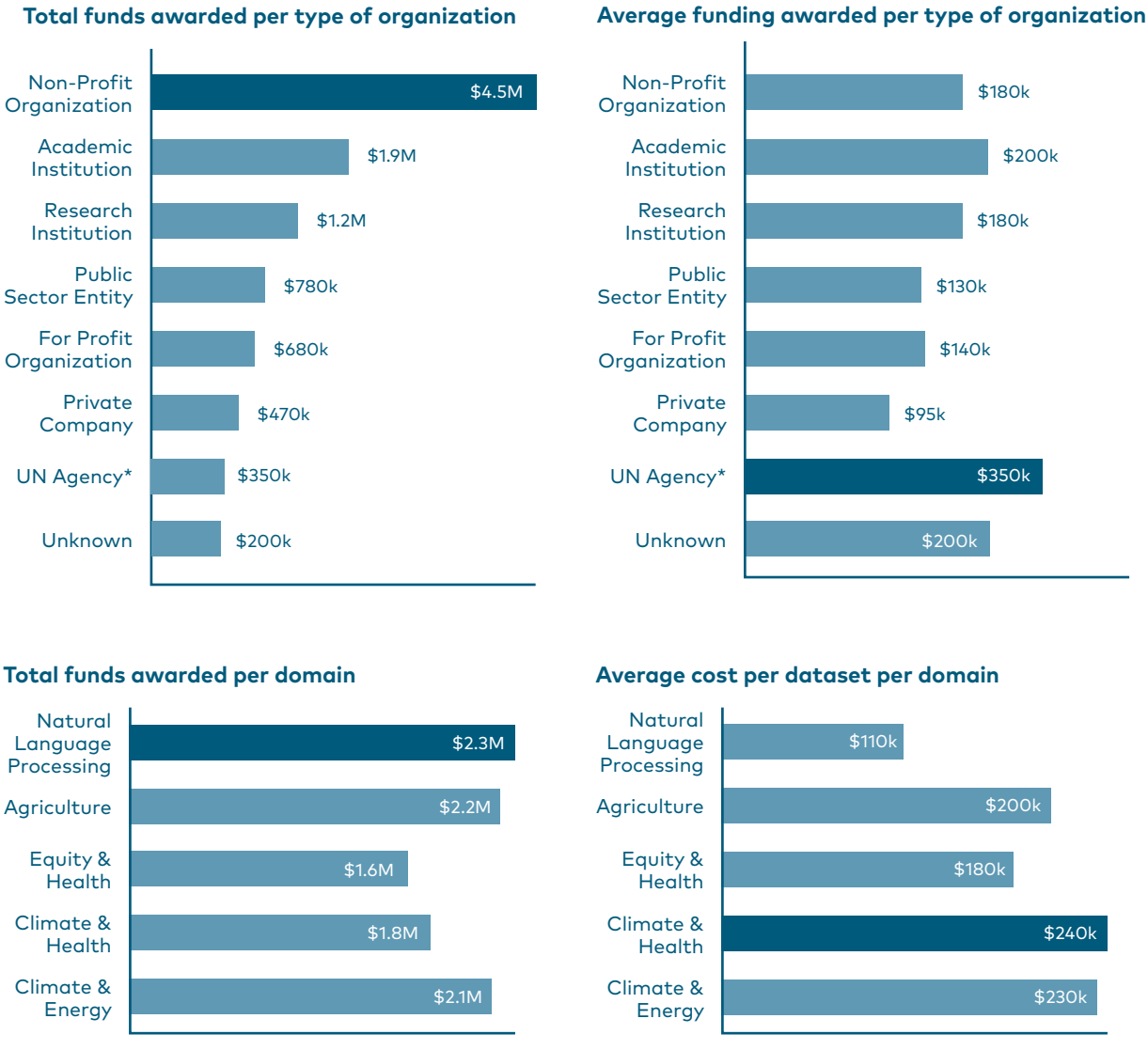


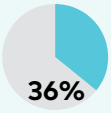
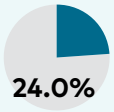
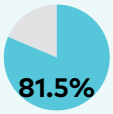
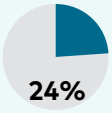
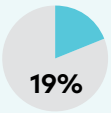
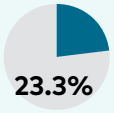
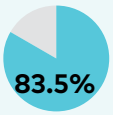
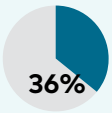
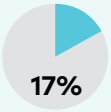
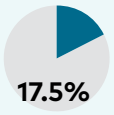
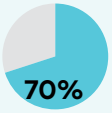
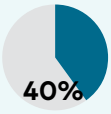
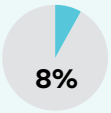
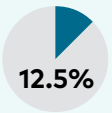
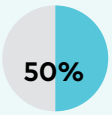
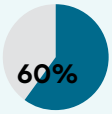
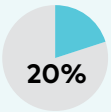
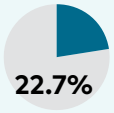
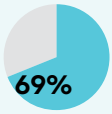
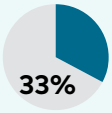
Figure 12: Total funds awarded per type of organization, and per domain

The average cost of funding a dataset ranges from **$107,000** to **$235,000**.



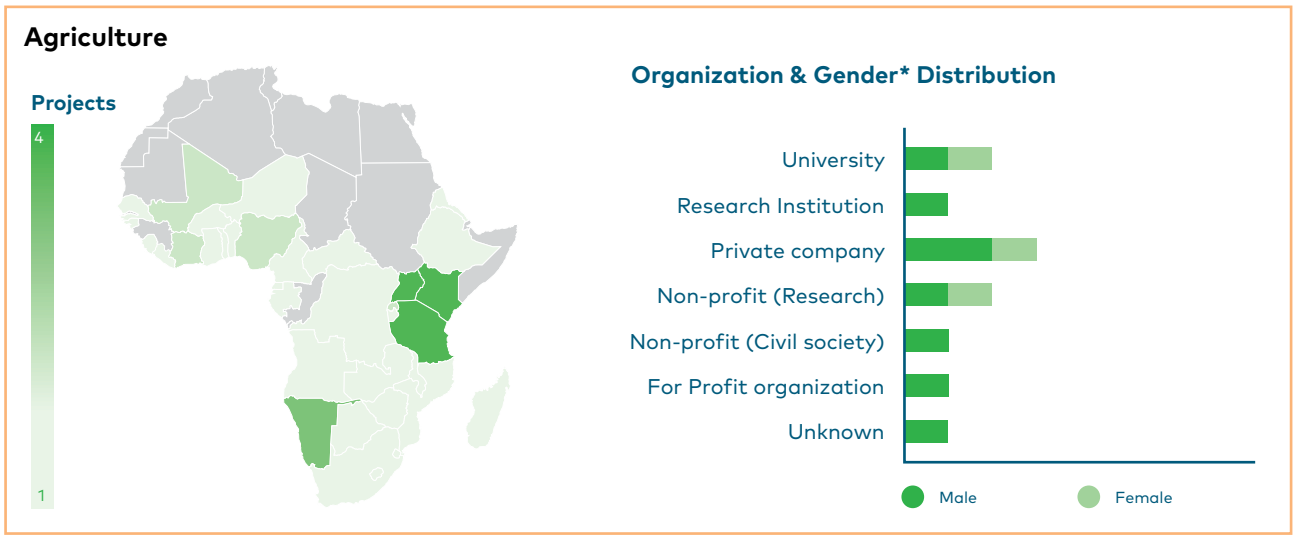
***Note:** The funding to this organization for this project is being distributed directly from one of Lacuna Fund's funding partners due to funder requirements

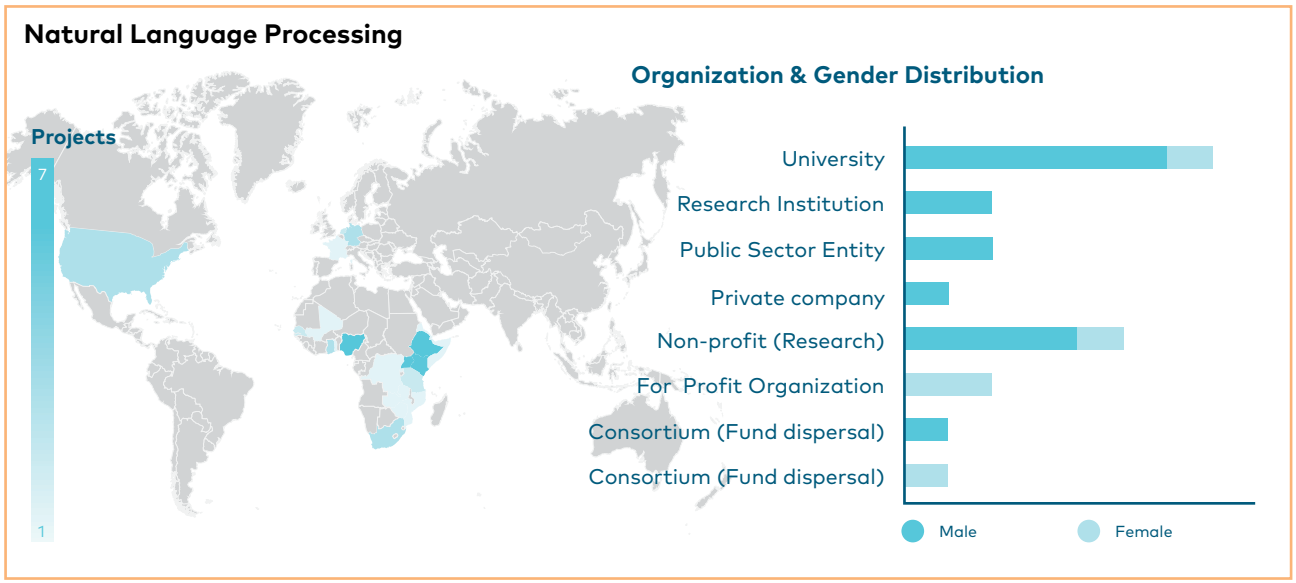
Table 4: Comparison of domains

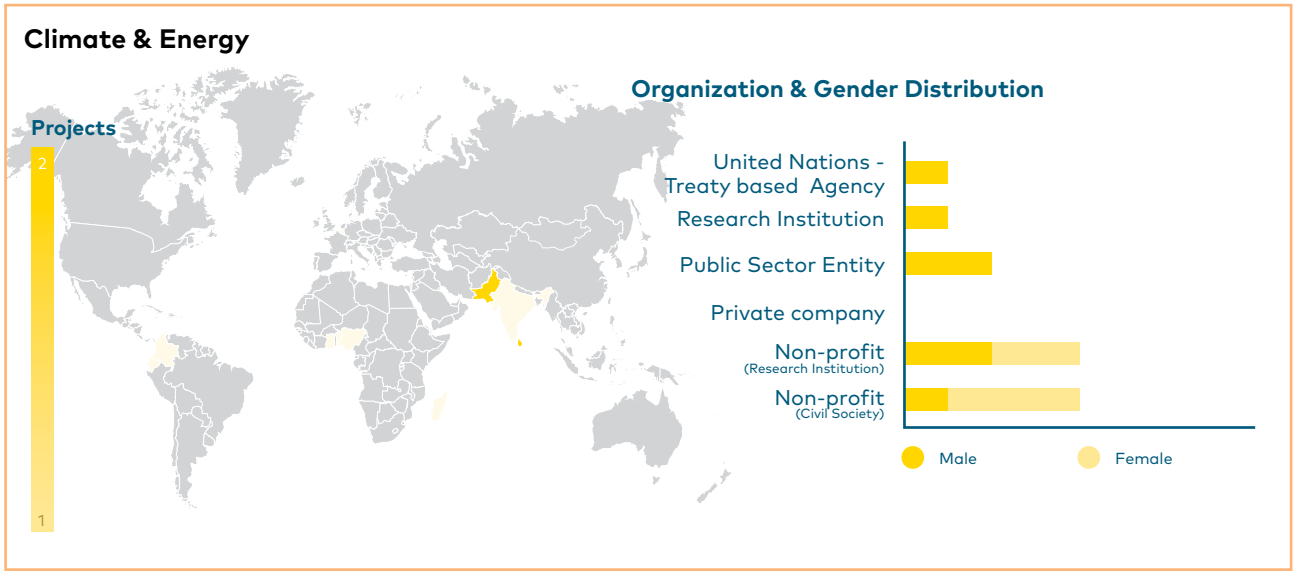
Domain	Year	#Projects funded	% of total projects	# Countries reached	Total funding awarded	% of total funds	% Project leads in LMICs*	%Female project leads
NLP	2020; 2021	21		31	$2.3M			
AG	2020; 2021	11		40	$2.2M			
E&H	2021	10		16	$1.6M			
C&H	2022	5		19	$1.2M			
C&E	2022	12		17	$2.1M			
Total	2020-2022	59	100%	54	$9.4M	100%	64%	59%

*Low and middle-income contexts

Figure 13: Comparison of domains







***Note:** Gender here refers to the gender of the lead applicant.

Climate & Health

Equity & Health

Relevant indicators

Indicators	Target	Actual
1. Number of applications received disaggregated by location of project, organization type	None yet	
2. Number of projects funded disaggregated by demographics; grant size; regions; research institutes.	None yet	59
3. Percentage of grants attributed to research groups based in low and middle-income contexts	None yet	100%
4. Number and value of "very close" declines	None yet	
5. Perceived level of effectiveness of process	None yet	75% of interviewed grantees (20) ranked quality of process high

PROCESS

3. What changes can Lacuna Fund make to its processes to continuously improve its ability to achieve its desired outcomes and influence?

In summary
Grantees shared that they would like better alignment between key dates and funding allocation, as well as receive frequent updates on requirements and useful resources to improve the grant process. In addition, they recommended that there be a platform that allows them to collaborate and share knowledge, this could include technology, social media and in-person engagement. Assessing whether datasets are accounting for biases in representation, inclusivity and other equity consideration should be incorporated into on-going monitoring, as well as assessing the impact of the datasets. Lacuna Fund has made improvements in processes, such as applicant webinars, extended project timelines where needed, and developing data quality guidelines. Overall, it was made apparent communication of changes and processes must be improved to support grantees awareness and engagement.

KEY DATA POINTS

Continuous learning support for grantees

- Lacuna Fund established some changes such as a catalog of resources on their website that is available to grantees and researchers at large. Some of the resource-related requests that emerged from grantees suggest that they may not be aware of this catalog, therefore more awareness-raising efforts to this regard are advised. Grantees requested more resources specific to all the relevant domains, and further understanding of the required quality metrics for sustainable datasets.
- Elements of further understanding such as these could be incorporated into Q&A webinars as those help during the application phase. Grantees also requested mentorship support, and other collaborative learning opportunities elaborated in the recommendations section. It is noted that capacity-building-initiatives have been incorporated, and it is suggested that these be made more visible and accessible to grantees through regular notification of learning opportunities on the grantees' new Slack platform and via email.

Communication post-grant award

- As represented in Figure 9, communication of TAP decisions, grant disbursement and project closure are areas of improvement needed in process. These are largely related to communication frequency and channels.
- TAP members highlighted that the disbursement of funds in cases where researchers collaborate with lead grantees in high-income countries should be investigated to ensure a fair distribution of funds within project. Allocation of additional fields in the budget and expenditure templates would assist in tracking this.

Project timelines

- The datasets for completed projects are available on Lacuna Fund website, and mostly 2021 and 2022 cohorts are still pending. A reason cited often for the delay in publishing datasets is that it takes several months just to collect the data. It is important to consider adjusting the time to completion and grant period to what the rate of completion so far dictates.

Measurement of dataset usage and outcomes

- Currently, the representation of marginalized communities can be best measured by assessing whether the datasets created are representative of people and geographies in LMICs or underserved communities in HICs. This has been found to be the case based on project locations and descriptions. Monitoring the population groups using the published datasets, as well as their intended purposes, would unlock further understanding of the impact of these datasets on machine learning models.
- Being able to track downloads is only possible on some hosting sites. Further knowledge in this area is needed and the guidelines produced by the Data Quality Advisor can support to this regard.

Relevant indicators

Indicators	Target	Actual
1. Reported level of changes that reflect grantee feedback - changes made to Lacuna Fund documents and processes reflect feedback from grantees and internal reflection	None yet	10 approaches to reported challenges identified

OUTCOMES

4. How and to what extent do complete datasets align with Lacuna Fund principles? To what extent do completed datasets address the equity or representation gaps they initially targeted?

In summary
Most grantees interviewed or surveyed are building new datasets as opposed to expanding existing datasets. 56% of proposals accepted said they would make an existing dataset more representative and inclusive, and at the time of evaluation it was found that this value increased to 78%. In addition, 13 out of 21 grantees surveyed reported that their project contributed to improved equity or better representation for underserved populations. When it comes to including marginalized groups in datasets, researchers are finding ways to operationalize fairness. The extent to which these commitments and alignment to principles are realized in complete datasets could not yet be rigorously assessed and should be considered in future evaluations once the appropriate processes to do so are established with grantees. An in-depth exploration of use cases beyond downloads is recommended following this evaluation.

KEY DATA POINTS

Datasets are addressing targeted representation gaps in ML

- All completed datasets in the Agriculture domain focused on food security challenges in Africa, which was the region of focus for the Agriculture RFPs in 2020 and 2021. This aligns well with Lacuna Fund's emphasis on creating datasets that are addressing gaps in datasets, inequity, and endeavoring to create transformational impact in underserved contexts. Similarly, NLP datasets focused on providing benefits to people in underserved contexts using NLP technology, particularly for people who speak under-represented languages.
- **78% of surveyed projects** are creating new high-value training crop or language datasets, which is an increase from the 56% indicated at the proposal stage. Specifically, 14 out of 20 of those interviewed, and 18 of the 21 grantees surveyed are building new datasets in mostly Agriculture and NLP domains. There are grantees who are both improving existing and developing new ones, indicating overall significant efforts to fill gaps and make available data more representative.

Open-source datasets for increased accessibility

- Sixteen of the 17 completed Lacuna Fund-funded projects so far have published datasets on open-source platforms, with the exception of one that was made to protect the privacy of the individuals in the data. Endeavors to ensure further accessibility include app development, student involvement, community engagement, leveraging existing resources and hosting the dataset in a public domain.
- One of the interviewed grantees mentioned that they purchased licensing or domain access for their datasets. Four grantees elaborated on the challenges they have faced which are regarding commercialization, copyright or privacy concerns.

Partnership and collaboration encouraging participatory research

- Grantees interviewed confirmed that they are partnering with local community members and leaders, contributors from government labs and ministries, local radio stations, hospitals, researchers, labelers, translators, technical institutions, innovation institutions, regional mapping centers, subject matter experts, agriculture forums and universities. These partners are mostly considered to be contributing technical expertise, project management capacity, and/or establishing access to reliable information and further funding opportunities.

Relevant indicators

Indicators	Target	Actual
1. Number of datasets created or expanded, by priority area (Health, Agriculture, Language, etc.) disaggregated by new/expanded/maintained	None yet	59 (21 NLP; 11 Ag; 10 E&H; 5 C&H; 12 C&E;)
2. Number and percentage of datasets published as open-source or appropriately accessible (if they cannot be open-source due to privacy laws)	None yet	16 datasets are open-source. 1 dataset is not open-source due to local privacy laws.
3. Number and percentage of datasets created or expanded that either: (a) make an existing dataset more representative and inclusive; (b) create a new, high-value training and evaluation dataset for an underserved population or under-represented crops or languages; or (c) makes an already widely used equitable dataset more sustainable		Grantee reporting a) 78% b) 70% c) 66% All grantees reported their datasets to satisfy at least one of the 3 requirements.
4. Expert views on field-level shifts, in the form of stories of changes in norms, are observed	None yet	7 new domains proposed; 6 trends affecting future of datasets reported

OUTCOMES

5. How and to what extent are Lacuna Fund-supported datasets being used, maintained, and updated to stay accurate and current? Who is using the datasets?

In summary
There was limited data on users of Lacuna dataset; an area for improved tracking going forward. However, grantees are also a user group, in cases where they develop models or other data use cases. Those that did use the datasets reported overall satisfaction with the datasets. Grantees are mostly focused on publishing datasets for now, as such do not report measures to ensure their datasets stay accurate and current. Datasets downloads provide additional indication of potential use. Those that have been downloaded from Lacuna Fund via website have been rated as average or good in quality. Many potential dataset users who responded to the survey were not keenly aware of Lacuna Fund's work but do have an interest in using the data, and indication for support in creating awareness of available datasets.

KEY DATA POINTS

Extent of use of Lacuna Fund datasets

- Grantees themselves are a main user group of Lacuna Fund datasets that provided the most insight to this evaluation question. Over half of the grantees interviewed are already modelling and maintaining the development of datasets, and almost half of the group interviewed have prioritized expanding existing datasets.

Profile of three dataset users:

- Two are academic researchers; One is a data scientist in the agriculture sector.
- One of the three have used the dataset to train a ML model to recognize maize plantation, two intend to use the data to research how to curb pollution through bioremediation, and to explore the effect of expansion of One Acre Fund on deforestation.
- The Agriculture dataset is intended to be used for modeling to benefit local citizens in the village, while the data downloaded to study deforestation is intended to benefit policymakers.
- Two of the three ranked the datasets as very relevant to their research, with the quality of the agriculture dataset rated as good, while that for the bioremediation research was reported to be average in quality.

- The vast number of known downloads of some of the existing datasets, which range between 705 to 312,000 downloads, which indicates that some of Lacuna Fund's datasets may be being used, maintained, or updated to a significant extent.
- Three out of 48 respondents to the data user survey downloaded datasets from Lacuna Fund website. This is not reflective of poor reach on Lacuna Fund's part, as it does not align with the large number of dataset downloads. An understanding of the market of dataset users is yet to be developed. Most of the respondents to the dataset user survey are researchers and program managers, and 80% of them are equally split between the non-profit sector and academia.

Domain focus of other dataset platforms

- Other download platforms mentioned as sources by respondents to the dataset user survey include Common Voice (Mozilla), FAO.org, GitHub, NASA, World Bank and UN. Datasets from these other sources, that are not necessarily funded by Lacuna Fund, are being downloaded to further research in the Agriculture, Education and Health sectors. Some of the challenges cited with these dataset platforms that Lacuna Fund could address to differentiate itself include the need for more linguistic detail and clarity, validation of reference data, better gender and indigenous representation.

Relevant indicators

Indicators	Target	Actual
1. Number/cases of instances of use	None yet	3 of 48 user survey respondents report use of a Lacuna Funded dataset
2. Percentage of funded projects that have sustainability plans	None yet	86% report having or developing sustainability plans
3. Percentage of projects that go on to get additional funding from other sources		10%

OUTCOMES

6. Who is likely to benefit from these applications? Or be left out of benefits or even harmed by these applications?

In summary
Multiple use cases are possible from datasets; this was observed by 65% of grantees interviewed, particularly in the Agriculture and NLP domains. The grantees themselves may benefit from the datasets in using them in model development or other applications. In addition, researchers and research institutes are reported users or potential users and have indicated the value in datasets in training. Other potential beneficiaries are community members that are exposed to the application of the datasets, e.g. farmers using crop monitors to detect disease, or patients who have more precise cancer or early stage diagnosis.

KEY DATA POINTS

Observed use cases

- Thirteen of the 20 grantees interviewed have observed more than one use case of their dataset, primarily due to the multiple applications that are possible with a dataset. Some unintended outcomes from the Agriculture domain include the improvement of community land use and the administration thereof through informed decision-making processes within government, which could potentially be an opportunity for another use case. Table 5 shares examples of other potential use cases of the datasets.

Table 5. Examples of reported potential use cases

Agriculture	Natural Language Processing	Climate & Energy
Disease classification Parasite identification Crop type mapping Livestock movement prediction Disease transmission	Automatic Speech Recognition (ASR) models Topic modeling Topic classification speech recognition Agenda bias evaluation	Predicting a community's electricity usage for better planning

In addition to the research being conducted by dataset users mentioned in the previous evaluation question, grantees interviewed foresee the following to be the impact areas of datasets being developed and modeled across the various domains:

1. **Improved access to information:** Data collected enabled the development of a mobile application that provides agri-advisory information to farmers. Initially available only in English, the application now includes an agri-advisory channel in Luganda, allowing farmers to ask questions (text and voice) and receive responses in their local language. This has increased accessibility and engagement, benefiting around 100 farmers in the pilot phase.

2. **Enhanced language technology:** The data collection efforts have contributed to advancements in language technology. Collaboration with organizations like Masakhane and Google has facilitated the development of state-of-the-art models for speech recognition and machine translation in various languages. The availability of a large volume of data has proven crucial in training effective models.

3. **Collaboration with large companies:** Recognizing the resource limitations faced by smaller teams, partnerships with large companies such as Google have been established. These collaborations involve sharing the collected data to improve existing models and provide feedback for model refinement. The aim is to leverage the expertise and resources of large companies to create more robust and effective language models.

4. **Adversarial testing and model improvement:** Adversarial testing, wherein attempts are made to break the models and identify weaknesses, has emerged as an effective strategy for model improvement. By subjecting the models to challenging edge cases, valuable feedback can be generated and incorporated into the training process, leading to rapid improvements.

5. **Empowering community experts:** The project aims to bridge the gap between community experts who understand language nuances and large institutions capable of developing advanced expertise in direct collaboration with the models. The goal is to create a synergy between human knowledge and machine capabilities, leading to better data and model integration.

6. **Real-world application and empathy:** The ultimate benefit of the entire process lies in the practical application of the developed models. By ensuring the models accurately understand and represent the nuances of the language, they can provide meaningful translations, avoid misrepresentation, and promote inclusivity. The focus is on building technology that genuinely serves and empathizes with the communities it aims to assist.

Relevant indicators

Indicators	Target	Actual
1. List of the communities or user groups that are prospective beneficiaries of the dataset applications	None yet	7 of 48 user survey respondents intend to use datasets, and 34 are not aware of the datasets

INFLUENCE

7. How and to what extent does Lacuna Fund's work influence other funders of datasets (whether NGOs, Companies, Governments or Philanthropies)? Are these investors making the work they fund public and are they considering dataset bias?

In summary
There are few other funders that have a particular focus on unbiased datasets, or those that exist are not easily found. The growth in funders partnering with Lacuna Fund, particularly philanthropic, is a signal of the influence the fund has had on expanding the pool of funders over the past three years.

KEY DATA POINTS

Collaborative funders

- Shifts in funding partners demonstrate trust and diversity amongst funders, depicted in Table 6.
- The big development players: USAID, GIZ, UNDP, Bill and Melinda Gates Foundation were shared as key influencers in the field.
- Apart from GIZ and Wellcome Trust, other 'big players' (as perceived by the bellwether interviewees) are not yet funders of Lacuna Fund or reported to be part of key events and meetings led by Lacuna Fund.
- Stakeholders have proposed increasing engagement with African philanthropic funds, those that focus on other points in the data value chain, to further promote the realization of change on the continent. This recommendation could be extended to Asia and Latin America.

Table 6: Funders at inception and current

Philanthropic		
The Rockefeller Foundation		
Gordon and Betty Moore Foundation		
Patrick J McGovern Foundation		
Robert Wood Johnson Foundation		
Wellcome Trust		
International development		
IRDC		
GIZ		
Tech company		
Google.org		

Other investors in unbiased dataset development

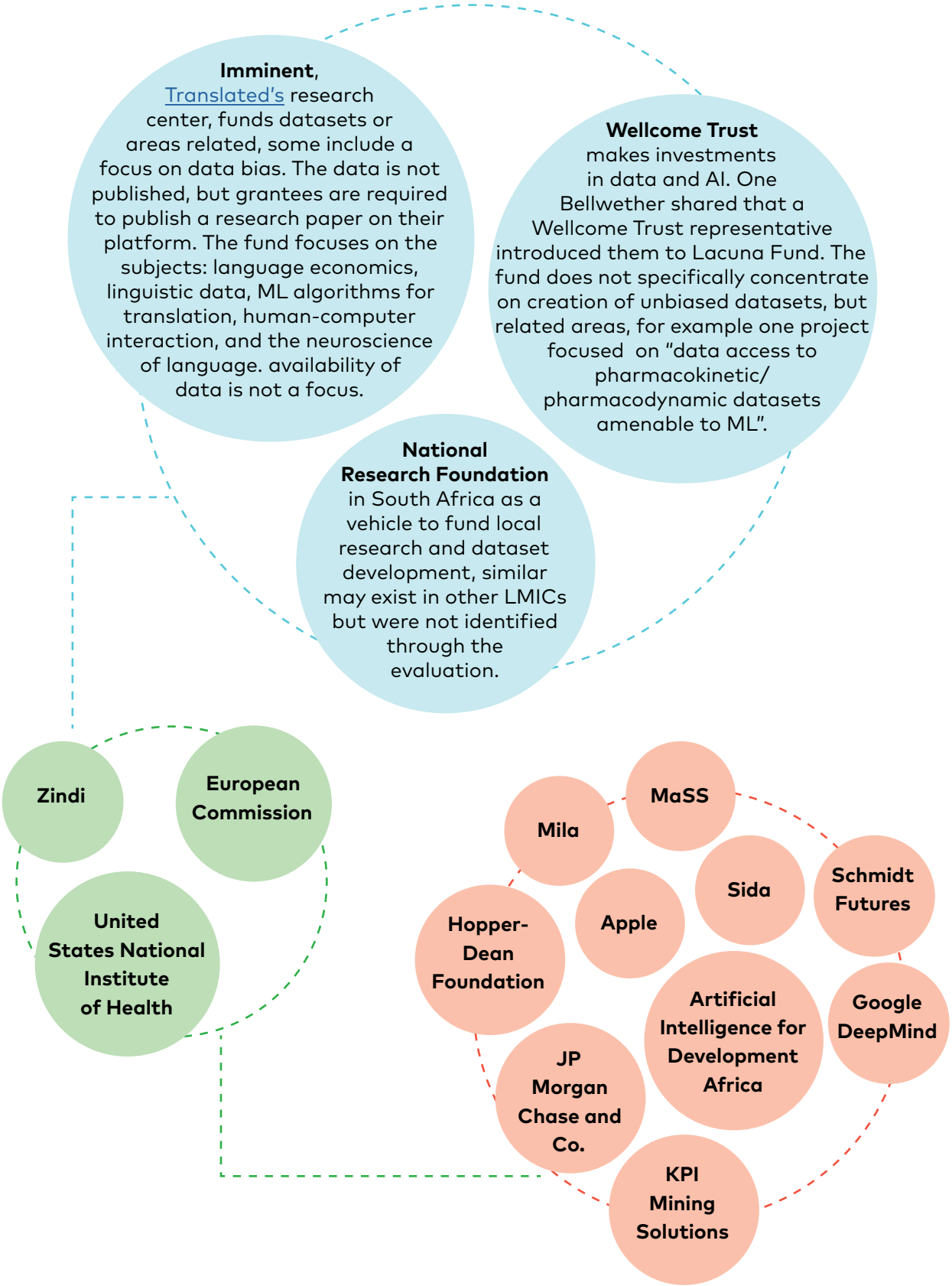
- Other funders of datasets, model development, open data, and related do exist. Few have an explicit focus on funding development of unbiased datasets and there is evidence of Lacuna Fund's role in influencing these funds. None indicate that the datasets are made publicly available but do share research on the data or using the datasets. Three funders were noted to be prominent in funding datasets, and others fund related areas or project types, as presented in the funder ecosystem diagram (Figure 14).

Figure 14: Funder ecosystem

Fund dataset development

Examples of funders across the value chain include funding model development, research related to ethical AI, education etc.

Funders of AI and ethical data events (e.g. AfricAI and Deep Learning Indaba) to take note of as potential funders of datasets



Relevant indicators

Indicators	Target	Actual
1. Perceptions and cases of changes in funder behaviour (directing funding to gaps, advancing open-source sharing requirements, etc.)	None yet	3 funders identified that fund datasets; more than 50 fund aspects of data value chain; demand for funding responsible AI and data increasing
2. Number of documented lessons shared on Lacuna Fund website	None yet	2 reports
3. Quantity and types of new funding coming into Lacuna Fund or being granted directly to researchers	None yet	4 funders in 2020 increased to 8 in 2021

INFLUENCE

8. How and to what extent do indirect benefits or unanticipated challenges accrue to the researchers funded by Lacuna Fund or to their research institutions (including to students who might work or train on the datasets) because of their involvement in the projects?

In summary
Grantees report indirect benefits because of Lacuna Fund, including personal growth, skills developed, visibility, additional funding and new projects attained by two grantees. Most grantees experienced challenges during their projects, mostly related to technical resources and unanticipated expenses. Lacuna Fund has made effort to address challenges, however, improvement is needed in communicating to grantees on changes or additions made to support them.

KEY DATA POINTS

Indirect benefits being realized by the researchers funded by Lacuna Fund

Indirect benefits included:

- influencing the initiation of a program, in turn accessing funding for AI related training/education (1)
- economic benefits from the models or data (4)
- improved stakeholder engagements and collaboration
- increased visibility and media coverage
- expanding networks
- Lacuna Fund has directly influenced access to additional funding for two additional Equity and Health applicants in 2021 from Wellcome Trust and Moore. Through the stimulus of Lacuna Fund, additional funding was also mobilized from Mozilla Common Voice and Fair Forward (GIZ and BMZ) for technical assistance initiatives in 2022 (sustainability plans and storage).

"[AFRICAI/MIKAI] are sponsoring our AI summer school, which is essentially training people on how to create models on that Lacuna data set, They are supporting us to have students come to Vancouver in Toronto this year and present their models. Lacuna Fund has opened up a floodgate."

- Grantee 20

BOX 1. GRANTEE INDIRECT BENEFITS
Personal Growth
Skills Development
Additional funding
Spin-off training and model initiatives
Visibility and media coverage

"...the NLP, the department, it is one of the reasons why in my university we are starting NLP as a program in our department, and because of that we are getting funding from France. France, there is this university in France that has a memorandum of understanding with the University of Nevada. They have promised to fund the diploma program at least I don't know the number of years, but at least until we start and run it for a couple of years."

- Grantee 12

Unanticipated challenges experienced by grantees

- Key challenges reported by grantees (18) are depicted in Figure 15. Other issues reported include: team capacity planning, unrest within communities affecting data collection, awareness of process and language barriers amongst annotators, recruitment processes, negotiating contracts and grants, managing partner delivery, time period of the project limiting, delays due to contracts and the expense of shipping equipment and bureaucracies and protocols of acquiring research permits and licenses.
- Resources and time were the more commonly reported challenge, which particularly refer to hardware and computing power and space, for example, one of the grantees shared "...Because you can imagine how long some of these health datasets have been collected for. I can give one example, a ward can have about 100,000 patients, followed up with for close to 30 years. That's a lot of data. Or if you had to combine the monthly climate indicators, the monthly rainfall for 30 years' worth of data in just one sub county in Tanzania, that's a lot of data. I think we kind of underestimated the computer power and cloud services that we needed.." Other resource related issues include the fact that six-month Google credits require repeat purchase for 12-month projects, and in some cases others are challenged by a shortage of computer devices.

Figure 15: Common Grantee Challenges

Lacuna has provided support to address some of the challenges through:

- Granting extensions to several grantees
- Offering in-kind cloud credits from Google.org
- Learning and reflection
  - Facilitated TAP reflection sessions to assess areas of learning and improvement
  - Annual Grantee Convenings to share and problem-solve challenges and lessons
- Asking grantees to outline a sustainability plan for their datasets
- Capability building
  - Webinars
  - Mentorship
  - Feedback
  - Workshops

Relevant indicators

Indicators	Target	Actual
1. Number of shared cases on additional benefits and extent reported by grantees (types)	None yet	6 grantees report additional funding for project or related initiatives; 6 report other areas of benefit
2. Quantity and type of conference presentations given or papers published by funded researchers	None yet	25 grantees report sharing research at conference, presentation and/or publication
3. No. of citations of funded researchers' work	None yet	3 grantees report a total of 18 citations
4. Extent to which grantees based at universities report student (graduate or undergraduate) engagement in the projects or access to the datasets that result	None yet	7 grantees reported engaging students in various activities

LANDSCAPE

9. What is changing in the context of Lacuna Fund's work?

In summary
Agriculture remains a key domain for dataset development, in addition to other emerging domains. Five trends in the context of Lacuna Fund's work include: increase in demand for ethical data for future scenario modeling; alternative business models for data access and sharing; generative AI; integrated knowledge for early warning; and graphical process unit hoarding. These need consideration in strategy and sustainability by determining the likelihood and impact on Lacuna Fund and grantees.

KEY DATA POINTS

The current context

- Limited data availability across sectors remains the status quo (Tadeusz Ciecierski-Holmes, Ritvij Singh, Miriam Axt, et. Al (2022)). This is acknowledged in Lacuna Fund Sustainability Plan, as the same issues Lacuna Fund intends to address through creation of datasets.
- However, already in 2021, the Secretariat and partners asserted the need to address issues such as interoperability, accessibility, and power dynamics which affect the usability of datasets that are created, expanded and maintained. The shifting landscape of researchers is evidenced in the growth of grassroot networks, such as those that developed open access data platform Open Streets (Solís and Zeballos (ed.) (2023)). These and the domains and trends to be described in this section require consideration through further research and deliberation in the next stage of Lacuna Fund.

Future domains

- The domains perceived as key to the future of Lacuna Fund include: Agriculture, Climate and Energy, Disaster, Engineering, Finances, Generative AI, Governance, Health, Language, Nutrition, Transport and Logistics. Agriculture remains the most emphasized and there are domains which Lacuna Fund currently funds and will remain relevant. Additionally, grantees shared cultural preservation, education, environment, online content and research as important future topics.

Bellwether reported trends

Trend 1. Increased use of models in future scenarios will create more demand for data

Anticipating and influencing the level of investment in new datasets is needed to ensure responsible data and the AI community are able to respond to the data demand. This may continue to require creation of unbiased datasets and potentially synthetic datasets.

Trend 2. Alternative business models to access and share data

Access to quality and robust data can come at high costs, yet sharing of datasets has limited economic benefit for grantees. 12 of 20 grantees share challenges related to open data which is an area of support needed, and accessing funders and stakeholders to influence. Influencing alternative models may support equitable cost and benefit of datasets amongst researchers. Tiered licensing or adaptations of Creative Common license that state conditions of use of data and creating data sharing agreement standards are emerging suggestions.

Trend 3. Generative AI proliferation

Generative AI global economy is anticipated to be valued at US$ 15 Trillion by 2023. There are still concerns around the potential risks

associated with generative AI, including deanonymization, model hallucinations and broadly unethical AI. Forums such as Responsible AI are discussing these topics. Debates and frameworks for responsible AI is an area where grantees and Lacuna Fund could play an influential role, and the availability of a knowledge sharing platform such as Slack can be used to encourage them to curate platforms to learn collaboratively and publish these learnings in their own right with their niche perspectives.

Trend 4. Integrated knowledge for early warning systems

Themes on integrating multiple datasets on different areas to understand complex issues are emerging, and the Climate related RFPs are exploring these themes. Grantee Convenings held by Lacuna Fund in 2022 and 2023 have allowed for interdisciplinary connections that can support complex modelling of this nature. An example is Health and Agriculture grantees discussing how monitoring livestock health could help predict and improve human health. These datasets may result in more advanced early warning systems or support preparedness for natural and man-made disasters. This may be a new domain to consider or projects to encourage, and Lacuna Fund could consider follow-on funding to develop multivariate models for related datasets.

Trend 5: Graphical Processing Unit (GPU) hoarding

Larger datasets and models require a significant amount of computing power. However, large tech companies have invested in purchasing graphical processing units which are required to store high quality data. Where GPUs are available they need consistent power and internet connectivity which is not the case in many low and middle-income contexts. Suggestions were made to explore innovative solutions to this challenge, such as supporting small data initiatives to grow and finding solutions that are not dependent on huge computing power.

> *"When a disaster happens we need to have rapid information and getting that collected using ML technology to rapidly get information and respond I think is another sector that is really not having enough information."*
>
> - Bellwether 1

LANDSCAPE

10. What strategies are helping grantees keep their datasets evergreen or sustain their usability in these ever-changing contexts?

In summary
Model building is the most common strategy to sustain usability of the datasets. Many grantees have developed sustainability plans. The detail and longevity of the plans was not assessed.

KEY DATA POINTS

Strategies to sustain usability: Model building

- Many grantees refer to building a model using the available dataset as a next step in their project. This is not specifically keeping the dataset evergreen, but focus placed on usability of the data. For example, understanding the effect of climate or extreme weather on maternal outcomes.

> *"For Amarik, we have already more than 60 head speech lexicons and we have collected more than 70 million Tweets. After filtering out, we have completed annotation of 10,000 Tweets. We have built the model on the data that we have collected in this research."*
> - Grantee 9

Strategies to sustain usability: Keeping existing datasets evergreen

- Some report that Lacuna Fund was an opportunity to keep existing datasets evergreen. The indirect benefits in visibility and receiving further funding are indicative of some opportunities to sustain usability of datasets through additional funds.

> *"Yeah, we're already building models in the data we had originally, which is why we needed more data. That Lacuna provided an opportunity for us, and that's what we're using to train our students."*
> - Grantee 7

Sustainability plans

- All grantees have considered sustainability approaches in proposal development and some have developed plans. The approaches include those listed below:

Table 7: List of sustainability approaches related by 21 surveyed grantees

Approach	Frequency
Leverage existing resources	● ● ●
App development	● ●
Hosting in public domain	● ●
Applied for funding	●
Student involvement	●
Cloud hosting	●
Community engagements	●
Incentivizing participants	●
Developing models to improve sales	●
Deployment of datasets	●
Frequent data collection	●

"With regard to sustainability, we have plans to create awareness about the dataset among NLP researchers and linguists interested in Igbo using various means such as social media, emails and conference presentations created for the purpose of publicizing the dataset..."

- Grantee 9

4.3 Key Lessons learned

The emerging outcomes and findings hold key lessons in overall implementation that have either supported or hindered achievement of outcomes and maintaining a principle-based approach. The below information provides the key insights which are aligned to the subsequent section on recommendations.

Deeper understanding of quality and participation

- The Data Quality Advisor was planning to review the quality of datasets at the time of this evaluation. The dataset users who reported downloading Lacuna Fund datasets rated them as average. It is important to measure the quality of generated datasets and to publicize the respective quality ratings, as is done by other dataset platforms, to help keep researchers informed. The work that has already started on dataset quality will support improvement in setting more clear standards for dataset quality.
- Participation and collaboration are highly valued by grantees. More coordination of the existing efforts and expansion thereof can stimulate even more cross-sectoral and grantee learning.

Tweaking grant processes

- The post-award communication processes are important to grantees and an area that needs improvement, such as a clarifying reason for rejection (if feasible) and guidance on support and communication around timeline changes.
- The 1-year project timeline is not sufficient for all project types and contexts – e.g. where experts need to be sourced or seasonal data needs to be collected.

Continued development of grantee skills

- Grantees value skills development initiatives and are particularly interested in enhancing presentation skills and research publications skills.
- Introducing capacity building initiatives and resources is valued by grantees and responsive to areas of need.

Responsiveness to common grantee challenges

- Common challenges experienced by grantees are varied, and not all can and will be anticipated. As described in the previous section, challenges include: technical resources and unanticipated expenses, data collection in unstable communities, awareness of process and language barriers among annotators, recruitment, contract negotiation, partner management, timeline constraints (with extensions granted), contract delays, shipping expenses, and research permit protocols.

Achieving transformational impact

- Transformational impact takes time. It requires a period of dataset uptake and usage, which is unlikely during the current grant period. Multiple use cases are possible and multiple users types can be expected.

Elevating awareness and influence of Lacuna Fund

- Lacuna Fund's influence on funders is growing and would benefit from more engagement events.
- Grantees are participating more at conferences and speaking engagements. They find value in collaboration and visibility through these events.
- Grantees and TAPs are key drivers of Lacuna Fund's vision and can be empowered to be drivers of engagement, learning and collaboration.

5. Recommendations

The findings collated in response to each of the evaluation questions provide sufficient lessons and insight to inform areas of improvement in fund processes, influence and relationships, sustainability planning, future model of operation, domains and opportunities. In alignment with the preceding section on lessons learned, the recommendations below are directed to Lacuna Fund Steering Committee, Secretariat, Funders and grantees, and these are listed below. Further prioritization may be needed to determine key recommendations to be actioned.

5.1 Recommendations for Lacuna Fund Steering Committee and Secretariat

Domain-related challenges being addressed by artificial intelligence

- Implement a **communication strategy** that reminds grantees of FAQs and other lines of support. Refine templates that will assist in streamlining implementation – e.g. more detailed budgets template to break down costs for planning purposes.
- Monitor common challenges closely and identify opportunities to support on a wide scale. This is included in the reporting template. An additional channel like a 'hotline' could be set up as a form or email address for ad hoc challenges that emerge during implementation. A focal communications lead can direct grantees to resources or flag for additional support
- Increase visibility and repeat **communication to grantees on multiple platforms** and languages to encourage participation. Regularly review skills development needs of grantees to ensure that capacity building is responsive. It may also be useful to collate a list of online and other opportunities that grantees can independently review.
- The function of Slack as a knowledge sharing and communication platform should be monitored to operationalize any prominent engagements. Additional regional or domain engagements is recommended, which could be lead by clusters of grantees, to strengthen grantee networks and collaboration.
- Improve the accessibility to TAP decisions for unsuccessful grantees to enhance their capacity to produce better proposals in future and broaden the reach of funds.

Understand the usage and impact of datasets beyond completion

- An in-depth **assessment of the impact of completed datasets** is recommended to evaluate the presence of any bias, the benefits to marginalized communities, and any potential harm. This would also aid in the development of fairness metrics. Ongoing modelling and use cases would make a good sample for this assessment. This may require agreement with grantees to engage them 1-2 years after implementation on use cases, and include long-term measurement of transformational impact within project budgets and plans.
- **Explore the use cases** through a combination of a needs assessment and an evaluation that can elaborate how transformational impact takes place and assess Lacuna Fund's contribution so that the results framework can be adapted accordingly. Grantees and other stakeholders within the network of Lacuna Fund are strong data sources for this metric, and can be incentivized in creative ways that help to scope the extent to which datasets are being downloaded and used
- Focused efforts are needed for **assessment of dataset quality**. The current Data Quality Advisor could assist to conduct in-depth data quality assessment of a sample of complete datasets, or be a resource for grantees to engage on data quality review during dataset development.
- Develop strategic communication and influence plan/strategy to support scaled fundraising efforts and increasing diversity in funders.

Enhance and expand the opportunities for outcomes to emerge

- Increase **visibility** of capacity building initiatives for grantees via email, newsletters and through Slack. Posters and other communiqué can also be sent to partner organizations and "Friends of Lacuna Fund" such as Masakhane, Data Science Africa, etc.
- Develop **guidelines and tools to support the development of use cases**. A participatory approach including grantees in the brainstorming process would help to clarify the required functionality of these guidelines.
- Institute a **Learning and Feedback system** where learnings from dataset development and relevant use cases are shared amongst grantees, with the goal to continuously improve underlying processes, activities and access to resources. In the hub and spoke model, the functionality of a learning and feedback system can be incorporated in a way that connects grantees to the Secretariat, grantees to one another, regions to one another, and the whole Lacuna Fund ecosystem to the landscape of ML.
- Research **alternative models for data sharing** to support grantees in diversifying funding and longer term economic benefit. Explore the case of malnutrition data in Chile as an example. Disseminate findings widely to influence future models.
- Maintain sustainability planning support and develop guidelines for future grantees.
- Support grantees' expressed willingness to self-govern localized collaborations/ gatherings for knowledge sharing and increased access to resources and support through funding, raising awareness and assisting to coordinate their efforts.

Increase awareness of the value of funding dataset development in low and middle-income contexts

- Consider an advocacy effort within the new model that drives conversation and guideline development for responsible AI practices with partners and other interested stakeholders in the landscape of ML datasets. As a grant-maker for public good, **awareness and advocacy** is needed to highlight the relevance of funding the development of datasets in under served contexts by indigenous and under-represented researchers.
- Thought leadership in responsible AI is a need in the broader data ecosystem. Grantees and hubs may be supported to publish lessons and **approaches to responsible AI** (in addition to dataset development).
- Develop detailed **events plan** with grantee contributions to track key events for participation. Develop a stakeholder map for key funders and other organizations to raise awareness of Lacuna Fund and grantees. Support Grantees and TAPs to establish hubs/groups for learning and collaboration which could scale awareness efforts. In addition, provide more social media guidance and tools for these stakeholders to engage on thought leadership topics tagging Lacuna Fund and partners.

Review and adapt the MEL framework to align with both the realized outcomes and the vision

- Institute a **monitoring and evaluation requirement for grantees'** accountability and continued learning and provide capability building in this area. Deliberate on whether this requirement is to be commissioned by the Secretariat or whether the responsibility could be placed with the grantees along with an associated budget within allocated funds.
- **Revise current results framework** to address whether some of the low-priority and medium-priority indicators should continue to be a part of the framework, and to consider the inclusion of indicators relating to other key findings from this evaluation.
- **Revise the Theory of Change** to update indicators with baseline values emerging from this evaluation, and to set informed targets for the short, medium and long term.
- Over and above the M&E framework requirements above, add a workstream

tracking tool that ensures a measurement and learning context that will enable the

Secretariat to continuously monitor, and track activities related to:

- LF principles and how these are being applied and are relevant to grantee support
- Unplanned activities that require immediate attention, and challenges
- New partnerships/ collaborations formed as a result of the projects
- Progress towards targets/ activities/ outcomes
- Downloads and subsequent usage of published datasets

- This tool will enable the routine monitoring and reporting of activities to stakeholders, and help to report on new activities that can contribute to their key indicators.
- **Track usage of the technical support resources** released on how to publish datasets, and treat this as a living document to enable further inclusion of other related challenges

5.2 Recommendations for Funders

Broaden funding collaboratives for further impact

- Expand the current funder list to increase the pool of **African funds**. Platforms like CAPSI may assist in identifying relevant philanthropists. The added diversity at the funder level would enhance understanding of fairness, equity and systems that perpetuate bias.
- Engage with key players in data sharing to **negotiate fair data access terms** (e.g. Twitter, Google) that enable further access to the landscape of ML for under-represented researchers.
- Consider expanding **funds for model building** efforts and identify stakeholders/partners supporting other stages of the data value chain to collaborate with hubs or form consortiums that support grantees to use and disseminate datasets.
- Consider adding to the funding focus the inclusion of existing grantees to use their already created datasets and pioneer use case development.

Deepen the reach of funds to accommodate more research efforts

- Consider funding the merging of similar thematic groups across domains to expand existing datasets and encourage more widely applicable use cases.
- Explore new programs to support and **build capacity of local researchers** (grantees), local entrepreneurs, and the communities to capitalize on the datasets and create value to society while having a sustainable business model.
- There are further use case opportunities to be realized from **domains emerging** from suggestions given by the sampled evaluation participants, these domains include Education, Disaster Management, Financial Service, Transport and Logistics. Current domains remain relevant.

5.3 Recommendations for Grantees

Coordinate efforts and collaborate for enhanced learning

- Form a Lacuna Fund Grantees **sustainability committee** that includes grantees and community stakeholders, possibly cutting across all domains, that are involved in creating datasets to share tools and optimize processes that serve the maintenance of datasets and continuation of research.
- Various partnerships have been established between grantees and **local organizations** that are contributing to the work of developing datasets. Explore these relationships for hub-specific roles and functions.
- Continue to **contribute further recommendations** at convenings and other engagements, with actionable insights that the Secretariat can engage and consider
- Initiate a framework that supports problem identification, data collection protocol identification, and platform development. Ensure that use cases from datasets are **individually mapped** to intellectual property and patented where necessary so their downloads can be tracked.

6. Conclusion

- The evaluation activities undertaken, along with the data collected and documents reviewed from all relevant stakeholders revealed a grant making entity that is making significant efforts to fund datasets for public good and ensure representation of under served and under-represented communities.
- Lacuna Fund's processes, from putting out a call for proposals, to making the unprecedented decisions of how to award funding for dataset development, to seeing grantees through to project close-out – are all well-structured and largely to the grantees' and TAP members' satisfaction. Areas for improvement have been noted in giving consistent support even post-award, developing platforms for collaboration and increasing access to resources, reporting templates, proposal feedback, etc. The participatory and inclusive approach set the tone for open communication and feedback and grantees are motivated, receptive and sufficiently impressed to recommend the fund to other researchers.
- There are clear opportunities to further this work and achieve the principle of impact through more interaction with the landscape of ML and responsible data practices that seek to discontinue biases subjected to marginalized groups and include them as significant population groups. Various domains have need for advances that can be made possible by additional ML datasets but the existing community of grantees could also be commissioned to pioneer the development of use cases in ways that continue to align with Lacuna Fund principles.
- This evaluation provides an opportunity to view the work of Lacuna Fund from different perspectives of different stakeholders, assess areas where improvements could be made and assess components of the Theory of Change that can be adapted to emerging results and key focus areas for the sustainability of the fund and its beneficiaries. The recommendations seek to strengthen the work of Lacuna Fund and establish a firm set of processes that lead to the desired outcomes and influence in the area of datasets for development.

7. Annexures

The data sources, sample selections, and instruments used throughout this evaluation can be found as listed below at the link supplied:

1. Theory of Change
2. Full Table of Indicators
3. Evaluation Sample Statistics
4. Data Collection Instruments
5. Data Sources

Link to the Annexures:
https://drive.google.com/file/d/1MGG_UQIysDnatGYDf2YEfk3ddKujrIAh/view?usp=drive_link


data
INNOVATORS


Lacuna Fund
Our voice in data