Expressions of Interest: Machine Learning Datasets for Language in Sub-Saharan Africa - 2021

Lacuna Fund: Our Voice on Data

19 October 2021

Table of Contents

1 – Introduction

Overview and Purpose of the Fund

Machine learning has shown great potential to revolutionize everything from how farmers increase their crop yields, to how governments communicate with their citizens during natural disasters, to how healthcare providers respond to global pandemics. But in low- and middle-income contexts globally, a lack of unbiased data puts these benefits out of reach. Lacuna Fund: Our Voice on Data is the world's first collaborative effort to directly address this problem.

Guided by machine learning professionals worldwide, Lacuna Fund provides data scientists, researchers, and social entrepreneurs with the resources they need to either produce new datasets to address an underserved population or problem, augment existing datasets to be more representative, or update old datasets to be more sustainable.

Since its launch in 2020, Lacuna Fund has worked to fill data gaps in several domains, supporting the creation, expansion, and maintenance of training and evaluation datasets in agriculture, health, and language.

Lacuna Fund aims to:

- Disburse funds to institutions to create, expand, and/or maintain datasets that fill gaps and reduce bias in labeled data used for machine learning.
- Make it possible for underserved populations to take advantage of advances offered by AI.
- Deepen understanding by the machine learning and philanthropy communities of how to most effectively and efficiently fund the development and maintenance of equitably labeled datasets.

Principles of the Fund

The following principles will guide the governance and operations of Lacuna Fund.

- **Accessibility** – Lacuna Fund is committed to ensuring that the datasets created through its funding are accessible to and benefit underserved communities in service of the goals outlined above. Datasets and related intellectual property will utilize appropriate open data licensing to maximize responsible downstream use. (see the Fund's IP Policy for additional details.)
- **Equity** – Lacuna Fund aims to make AI more equitable by supporting datasets that are created by and responsive to the needs of those with underrepresented identities globally. These datasets should not create or reinforce bias (e.g., they should be gender inclusive and representative of people of color globally), nor should they support systems or technologies that create harm.

- **Ethics –** Lacuna Fund will fund data collection in a manner consistent with ethical labor standards and require recipients to specify steps they will take to protect privacy and prevent harm in the collection, licensing, and use of datasets created with grant funds.
- **Participatory Approach –** Lacuna Fund strives to meet the needs of affected stakeholders by encouraging the leadership or strong engagement of local experts, beneficiaries, and end users in the governance of the Fund and in supported projects. The Fund will consider participation in a manner consistent with our principles on equity and ethics.
- **Quality** – Data generated by Lacuna-funded efforts should be of high quality, enabling beneficial applications in research, communities, and industry.
- **Transformational Impact** – Lacuna Fund aims to unlock the advances offered by AI for poor and underserved communities by funding datasets that address fundamental gaps in AI.

Philosophy of Grantmaking

Lacuna Fund values a collaborative and locally driven approach to data creation, expansion, and maintenance. We recognize that the continued usefulness and maintenance of open data derives from a community invested in that data.

Lacuna Fund hopes to fund datasets that contribute to multiple applications of high social value, whether through research, commercial innovation, or improved public sector services. While "Section 3: Purpose and Need" sets out needs identified by the Technical Advisory Panel (TAP), Lacuna Fund welcomes novel ideas within the domain area that have a clearly articulated benefit aligned with the selection criteria listed below.

This call regarding datasets for African NLP is supported by The Rockefeller Foundation, Google.org, Canada's International Development Research Centre, and GIZ on behalf of the German Federal Ministry for Economic Cooperation and Development.

2- Overview

Organizational Eligibility

Lacuna Fund aims to make its funding accessible to as many organizations as possible in the AI for social good space and cultivate capacity and emerging organizations in the field.

To be eligible for funding, organizations must:

- Be either a non-profit entity, research institution, for-profit social enterprise, or a team of such organizations. Individuals must apply through an institutional sponsor. Partnerships are strongly encouraged; however, only the lead applicant will receive funds.
- Have a mission supporting societal good, broadly defined.
- Be headquartered in or have a substantial partnership in sub-Saharan Africa.
- Have all necessary national or other approvals to conduct proposed research, as well as data use agreements or plans to secure them. The approval process may be conducted in

parallel with grant application, if necessary. Approval costs, if any, are the responsibility of the applicant.

- Have the technical capacity – or the ability to build this capacity through a partnership described in the EOI - to conduct dataset labeling, creation, aggregation, expansion, and/or maintenance, including the ability to apply best practice and established standards in the specific domain (e.g. natural language processing) to allow high quality AI/ML analytics to be performed by multiple entities.

Selection Process and Evaluation Criteria

Lacuna Fund seeks Expressions of Interest (EOIs) from organizations that are interested in unlocking, creating, aggregating, and/or improving training and evaluation datasets that can support natural language processing for languages in sub-Saharan Africa. Lacuna Fund and its partners will perform an initial screen of EOIs for organizational eligibility and feasibility. Technical Advisory Panel members may not submit an EOI or a proposal in response to an RFP for which they are a reviewer (see Lacuna Fund's Conflict of Interest Policy).

The Technical Advisory Panel for this call will assess EOIs to determine a short list of organizations that will then be invited to provide full proposals for funding. Selections for the short list of invited applicants will be based on the degree to which they meet the following criteria:

- **Quality –** The organization or team proposing the project demonstrates interdisciplinary collaboration between qualified experts in language, research, machine learning, and data management. Partnerships between well-resourced actors with strong research capacity and actors serving marginalized populations and low-resourced languages are encouraged.
- **Transformational Impact –** Datasets should improve machine learning and imply use cases that enable a demonstrable social benefit for underserved populations. Examples include, but are not limited to: a) labeling, cleaning, collecting validation data for, or pooling existing datasets to unlock additional value or ensure greater accuracy in the existing dataset; b) creating a new, high-value labeled dataset for a low-resourced language; c) making an existing dataset more representative and inclusive of race, gender, ethnicity, ability, etc.
- **Equity –** There is a compelling theory of change demonstrating how the dataset will be applied to help vulnerable and underserved communities.
- **Participatory Approach –** Datasets should center the needs of affected communities and work with partners to identify community benefit. Specifically, projects should engage community members in data governance decisions, (e.g., what data is curated and how it is used). If the dataset has a geographical scope, the team is predominantly located in the respective area and/or sustains close ties to local actors to ensure sustained maintenance and usage of the dataset by the local community.
- **Ethics –** The project is able to pass an ethical screen (e.g., an institutional review board) that probes: a) privacy concerns, b) potential for downstream misuse c) possible discrimination vectors (e.g., gender), and d) fair and equitable working conditions, if paid labelers are involved in the project.

- **Efficiency –** The proponent has considered existing datasets and proposes to use effective data collection and labeling techniques and tools to speed the collection, cleaning, and sharing of data.
- **Feasibility –** The project is feasible in relation to the budget and scope of work proposed. While the EOI is not expected to include a detailed budget, we do expect EOIs to include a broad estimate of the overall budget expected to complete the project.
- **Accessibility -** The dataset will be made widely accessible under open-source licensing pursuant to Lacuna Fund's IP Policy. If this is not possible, a compelling case is made for more restrictive licensing in order to protect privacy or prevent harm, along with a mechanism for providing access under the proposed licensing.
- **Sustainability –** The project has a plan to ensure sustainability and future maintenance of the dataset, e.g., by a dedicated community or a pool of interested parties (for-profit and/or not-for-profit) and a robust governance model for the open dataset.

Timeline

EOI Call Posted Publicly on Lacuna Fund Website	**19 October 2021**
Question and Answer Deadline Please submit questions to secretariat@lacunafund.org	**2 November 2021**
Answers Posted	**9 November 2021**
Expressions of Interest Due	**1 December 2021**

Question and Answer Period: All questions related to the EOI should be submitted to secretariat@lacunafund.org with "Language 2021 Question" in the subject line. Questions submitted by 5 November 2021 will be de-identified and answered publicly by 15 November 2021 on the Lacuna Fund website in a document posted on the "Apply" page.

3 – Purpose and Need

Purpose

The purpose of this call for EOI is to identify projects to submit full proposals to develop open and accessible datasets for machine learning applications that will enable natural language processing for languages in sub-Saharan Africa. The ability to communicate and be understood in one's own language is fundamental to digital and societal inclusion. Natural language processing techniques have enabled critical AI applications that facilitate digital inclusion and improvements in numerous fields, including: education, finance, healthcare, agriculture, communication, and disaster response, among others. Many advances in both fundamental and applied NLP have stemmed from openly licensed and publicly available datasets.

However, such open, publicly available datasets are scarce to non-existent for many African languages, and this means the benefits of NLP are not accessible to speakers of these languages. Where relevant datasets do exist, they are often based on religious, missionary, or judiciary texts, leading to outmoded language and bias. There is a need for openly accessible text, speech, and other datasets to facilitate breakthroughs based on NLP technologies for African languages.

Need

Lacuna Fund seeks Expressions of Interest (EOIs) from qualified organizations to develop open and accessible training and evaluation datasets for ML applications for NLP in sub-Saharan Africa. The TAP recognizes the importance of datasets that would create significant impact regardless of the number of speakers of the included language, as well as the need for multi-lingual datasets.

EOIs may include, but not limited to:

- Collecting and/or annotating new data;
- Annotating or releasing existing data;
- Augmentation of existing datasets in all areas to decrease bias (such as gender bias or other types of bias or discrimination) or increase the usability of NLP technology in low- and middle-income contexts;
- Creating small, higher-quality benchmark data for NLP tasks in low-resource African languages.

While the focus of Lacuna Fund is primarily on dataset creation, annotation, augmentation, and maintenance, **proposals may include the development of a baseline model** to ensure the quality of the funded dataset and/or to facilitate the use of dataset for socially beneficial applications.

The TAP sees a need for datasets that enable better execution of core NLP tasks in African languages, as well as the assessment of systems performance in African languages, **including but not limited to the following**:

- **Speech corpora,** including for applications that allow illiterate or otherwise underprivileged groups to access technology tools, information, and/or services. In particular, there is a need for the creation of large, high-quality speech corpora (i.e., representative of population, phonetically rich and balanced, clean spelling, accurate orthographic transcriptions, no paraphrase or removal of disfluencies) or datasets aimed at supporting textless NLP systems.
- **Labeled text corpora** for use as training or benchmark evaluation data, including parallel corpora for machine translation or corpora to support other fundamental or downstream NLP tasks. Downstream tasks might include, but are not limited to: question answering and conversational AI, sentiment analysis datasets; automatic text summarization or other natural language understanding and generation tasks, or resources to support NLP education.

- **Unlabeled text corpora** for language models that support multiple avenues of research or application. This includes text corpora that can be used to support the training and evaluation of speech models.
- **Datasets related to code-switched text or speech** that improve the performance of NLP tasks in such situations.
- **Domain-specific creation or augmentation of text and speech datasets**, such as digit datasets, place names, or specific word pairs or sentences, that enable applications with significant social impact.
- **Multimodal and other innovative datasets**, such as video or audio captioning or other image-text interactions.

Across all proposed projects, the TAP encourages teams to consider:

- Leveraging existing resources, including collection methods or tools as well as existing corpora or models in higher-resourced languages.
- Strong collaboration and contribution from linguists and AI ethics researchers on project teams.
- Sociological elements of NLP (e.g., differences in verbally describing the same picture across cultures).
- Expanding past most popular/populous languages to grow the community.
- Including regional and other variations in datasets.

Review projects selected in our 2020 round of funding here to see what work is currently underway. We welcome projects that take completely new approaches and address different languages or build upon and gain inspiration from these supported projects.

4 – Expression of Interest Information

Expressions of Interest will only be accepted through the application portal available at www.lacunafund.org/apply. A description of application questions is available below for information only. ***Please limit your expression of interest to 4 pages not including references**, with 2.5 cm margins and a minimum of 11-point font. Appendices or proposal narrative material beyond 4 pages will not be reviewed.*

The following information is required:

- **Descriptions and Qualifications of Participating Project Teams** – Provide brief background on the mission of participating organizations, services provided, and constituencies served; how they satisfy the eligibility criteria articulated above; and the applicant's unique qualifications to undertake the proposed work, including experience developing and sharing language datasets.
- **Project Description and Specifications** - Please briefly summarize the dataset you intend to create, augment, or maintain, and the specific machine learning need the dataset would fill within the context of the particular language(s) or NLP task(s). If possible, please provide

specific information about the scale (e.g. hours or sentences of data) and contents (e.g. format and contents of labels if applicable) of the proposed dataset.

- **Anticipated Use Cases and Benefits** – What relevant use cases will the dataset enable? Please describe the social and cultural context for the potential resultant language technologies, and if applicable, describe how the dataset, models and related products could motivate and enable multiple and durable paths of application. The Technical Advisory Panel is interested in datasets that further the state of research and practice in African NLP as well as those that enable socially beneficial applications.
- **Project Methodology** – Explain how data will be collected, labeled, or augmented. Describe steps the project will take to ensure representation and avoid bias, as well as ensure quality in the collected or labeled data. If data is intended to be crowdsourced, please provide information on existing engaged communities, incentives, or the feasibility of the approach.
- **IP, Ethics, and Privacy** – Please confirm that the proposed dataset and related intellectual property will be licensed under a CC-BY 4.0 International license per Lacuna Fund's IP Policy. If more restrictive licensing is proposed, please provide a clear rationale. If the project intends to use or aggregate existing datasets, corpora, or other resources, please provide information on availability and licensing.
- **General Timeframe and Overall Budget for Project Implementation**:
    - Timeframe – Share a broad timeframe for completion of the steps included in the General Methodology above, including total number of months required for completion.
    - Budget – Provide a broad overview of the expected budget for completion of the steps included in the General Methodology above. Budgets should be submitted in US Dollars. The total pool available is approximately $900,000 USD. We are anticipating proposed budgets in the range of $10k – 100k for small to medium-sized projects and up to $200k for large, complex projects. We anticipate being able to fund 2-3 large projects and a larger number of smaller projects. **Indirect costs are limited to 12% of the budget.**

Thank you for your interest in Lacuna Fund and your efforts to make machine learning applications in the field of natural language processing. We look forward to reviewing your submission.