Lacuna Fund
Our voice on data

Lacuna Fund: Our Voice on Data

Request for Proposals: Datasets for Language in Sub-Saharan Africa

23 September 2020

1 – Introduction

This request for proposals (RFP) will fund the creation, labeling, augmentation, or maintenance of datasets for machine learning in languages in sub-Saharan Africa. We envision that these datasets will both enable specific tasks in natural language processing (NLP) and broader research in machine learning with the ultimate goal of supporting social impact.

This RFP closes on 6 November 2020. Proposals will only be accepted through the online portal on Lacuna Fund's website (http://www.lacunafund.org/apply). This document contains further details about the RFP's eligibility criteria, selection criteria, timeline, Q&A process, and purpose and need, as well as application requirements.

Thank you for your interest in Lacuna Fund, and for your dedication to closing data gaps to allow the machine learning community to better solve urgent problems. We look forward to receiving your proposal!

Purpose and Goals of the Fund

Lacuna Fund supports the creation, expansion, and maintenance of equitably labeled* open datasets that enable the robust and more equitable application of machine learning tools of high social value in low- and middle-income contexts globally.

The 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, development, and philanthropy communities of how to most effectively and efficiently fund development and maintenance of equitably labeled datasets.

** Both labeled and unlabeled training and evaluation data is critical to advancing the state of NLP for underserved languages. This RFP supports the creation, augmentation, and maintenance of training and evaluation datasets that serve the goals outlined in Section 3: Purpose and Need.*

Principles of the Fund

The following principles guide the operations of Lacuna Fund.

- **Accessibility –** The Fund is committed to ensuring that labeled datasets created through its funding are openly 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 –**The Fund aims to make AI more equitable by creating datasets that are representative of the Global South and people of color globally and their needs. These datasets should not create or reinforce bias.
- **Ethics –** The Fund will fund data collection in a manner consistent with ethical labor standards and requires subgrantees to outline steps they will take to protect privacy and prevent harm in the collection, licensing, and use of datasets created with grant funds.
- **Participatory Approach –** The Fund strives to meet the needs of affected stakeholders by involving local developers, researchers, beneficiaries and end users in the governance of the project as well as in data creation.
- **Quality –** Data generated by Lacuna-funded efforts should be of high quality, enabling beneficial applications in research, communities, and industry.
- **Transformational Impact –** The 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.**

Lacuna Fund is supported by The Rockefeller Foundation, Google.org, Canada's International Development Research Centre, and the German development agency GIZ on behalf of the Federal Ministry for Economic Cooperation and Development (BMZ).

2- Overview

Organizational Eligibility

The 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 a non-profit entity, research institution, for-profit enterprise, or a team of such organizations. Individuals must apply through an institutional sponsor. Partnerships are encouraged, but only the lead applicant will directly receive funds.
- Have a mission supporting societal good, broadly defined.
- Be headquartered in Africa or have a substantial partnership with organization(s) headquartered in Africa.
- Have all necessary national or other approvals to conduct proposed research. The approval process may be conducted in parallel with grant proposal, if necessary. Approval costs, if any, are the responsibility of the applicant.
- Have the technical capacity to conduct dataset labeling, creation, expansion, and/or maintenance, including the ability to apply best practice and established standards in the specific domain (e.g. NLP) so that multiple entities can utilize datasets to create high quality AI/ML outputs.

Selection Process and Evaluation Criteria

The Lacuna Fund and partners will perform an initial screen for organizational eligibility and completion of all proposal requirements. Following the initial screen, a Technical Advisory Panel of domain experts, data users, and stakeholders will evaluate proposals based on the selection criteria outlined below. Technical Advisory Panel members may not submit a proposal in response to an RFP for which they are a reviewer (see Lacuna Fund's Conflict of Interest Policy).

- **Quality –** The organization or team proposing the project includes qualified experts in: a) machine learning; b) data management and data licensing; and c) if applicable, the domain of interest;
- **Transformational Impact –** The project either: a) unlocks additional value in an existing dataset; b) creates a new, high-value training dataset for an underserved population or problem related to the Sustainable Development Goals (SDGs); c) makes an existing dataset more representative and inclusive of low- and middle-income contexts; or d) makes a widely used and equitable dataset more sustainable.
- **Equity –** There is a compelling theory of change demonstrating how the dataset will improve machine learning and be applied to help the poor, vulnerable, or other populations underserved by current language technology.
- **Participatory Approach –** If the dataset has a geographical scope (e.g. language or geospatial datasets), 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. (See ACM's Code of Ethics for general standards.)

- **Efficiency –** The proponent has taken into account existing datasets and proposes to use effective data collection and labeling techniques and tools to speed the collection, cleaning, maintenance, and sharing of data.

- **Feasibility –** The project is feasible in relation to the budget and scope of work proposed.

- **Accessibility –** The dataset will be made widely accessible under open licensing, or if this is not possible, a compelling case is made for more restrictive licensing in order to protect privacy or prevent harm. This includes a plan on how to ensure that the data source from which the dataset is created does not fall under a license or copyright that prevents making the dataset widely accessible under open 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

Timeline

RFP Posted Publicly on Lacuna Fund Website	23 September 2020
Question and Answer Deadline Please submit questions to secretariat@lacunafund.org	7 October 2020
Anonymized Answers Posted Publicly on Lacuna Fund Website	14 October 2020
Applications Due	6 November 2020

Question and Answer Period: All questions related to the RFP should be submitted to secretariat@lacunafund.org with "Language RFP 2020 Question" in the subject line. Questions submitted by 7 October will be de-identified and answered publicly by 14 October on the Lacuna Fund website in a document posted on the "Apply" page.

3 - Purpose and Need

This request for proposals will fund labeled or unlabeled datasets to enable novel and improved work in natural language processing (NLP) for languages in sub-Saharan Africa. Lacuna Fund's mission is to

support the creation, labeling, expansion, and maintenance of transformative datasets to fill gaps and improve accuracy in machine learning.

Purpose

The ability to communicate and be understood in one's own language is a fundamental right and a prerequisite to digital and societal inclusion. Natural language processing (NLP) techniques have enabled critical applications to achieve this—to improve education, financial inclusion, healthcare, agriculture, communication, and disaster response, among many other areas.

However, publicly available datasets are scarce to non-existent for most African languages.[1] Where these datasets do exist, they are often based on religious, missionary, or judiciary texts, leading to outmoded language and bias. Across text, speech, and other datasets, the gap in openly accessible datasets has prevented breakthroughs based on NLP technologies. Labeled data and speech corpora remain a key element of this gap, as well as the availability of corpora that can be used in transfer learning or semi-supervised approaches.

This RFP will begin to fill this need for language data, including both labeled data and corpora that fill other key data gaps. Notably, Lacuna Fund's efforts in NLP build on a recent groundswell of momentum to create better and more open NLP tools in African languages from ML community members, including recent academic workshops, volunteer collaborations, innovative academic programs, and other efforts.

Need

This RFP aims to broadly fund open training and evaluation datasets for NLP in African languages. The TAP recognizes the diversity and importance of NLP needs in African languages, as well as the need for multilingual datasets. Proposals should move forward the current state of data and potential for the development of NLP tools in the language(s) for which efforts are proposed.

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

- Speech corpora, particularly enabling automated voice recognition that allows illiterate or otherwise underprivileged groups of persons to access information and/or services;
- Labeled and unlabeled text corpora for use as training data;
- Parallel corpora for machine translation;
- Corpora to support fundamental NLP tasks, such as named entity recognition (NER), part of speech tagging, embeddings, etc.;
- Datasets for key downstream NLP tasks, such as question answering and conversational AI, sentiment analysis datasets, or technology for language education;
- Datasets to improve the performance of NLP tasks on code-switched text or speech.

[1] https://arxiv.org/pdf/2004.09095.pdf

More broadly, there is also a need for:

- 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;
- More benchmark data for NLP tasks in underserved languages or to inform multilingual models;
- Innovative datasets, such as video or audio captioning or other image-text interactions.
- 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.
    - As an illustrative example, a potential application might include radio keyword spotting for humanitarian response, speech recognition for agricultural inputs, or another need that requires a specific dataset.

Note that this RFP solely supports the creation, expansion, and maintenance of training and evaluation data for machine learning. While this includes support for semi-supervised and active learning approaches to the collection or generation of data, Lacuna Fund is unable to support model or application development.

While the Technical Advisory Panel has set out some key needs in the domain above, applicants should feel free to propose novel ideas within the domain outlined in the "Purpose" section and aligned with the evaluation criteria of this RFP (See "Philosophy of Grantmaking" on page 3 for further details).

All proposals should:

- Consider how to ensure collected data is representative and to the extent possible, unbiased.
- Ensure compliance with and specify relevant copyright or IP law for any source text.
- Use existing or common infrastructure (see the "Resources" page on the Lacuna Fund website for a non-exhaustive list).
- Consider innovative and cost-effective means to collect and maintain quality data, including community-based collection, commons-based peer production, other forms of crowdsourcing, gamification, and the use of active learning techniques.
- Demonstrate awareness of and attempt to be interoperable with related, existing datasets.
- Ensure your metadata is well-described. (see "Data Statements for Natural Language Processing" for further information on best practices. A data statement is NOT required for this proposal.)

4 – Proposal Information

NOTE: The Lacuna Fund website includes various resources, such as relevant references on data quality, documentation, and format, to help applicants prepare a competitive proposal.

Proposal submissions 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. **The following sections include:**

- Applicant Information (accessible in the portal)
- Proposal Narrative
- Letters of Support (optional, required if use of a non-public source dataset is proposed)
- Timeline, Budget, and Budget Narrative
- Due Diligence Self-Reporting Form

Your Information

This section will prompt the applicant to provide:

- A 200-250 word proposal abstract;
- Details about the institution(s) and/or team applying;
- Where the work will take place;
- CVs for key team members.

In the future, the applicant will be asked to provide information about:

- The affiliated institution(s) ethical review processes;
- The team's ability to gain national approvals.

Proposal Narrative

Please limit your proposal narrative to 10 pages not including references, with 2.5 cm margins and a minimum of 11-point font. Appendices or proposal narrative material beyond 10 pages may not be reviewed.

This section will prompt the applicant to upload a cohesive narrative, in PDF or Word format, that addresses the following:

Qualifications - Describe the organization(s) or partnership(s) applying and your unique qualifications to undertake the proposed work.

Proposed Dataset and Use Cases: 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).

> *Would your proposed dataset, for example, enable others to build a part of speech tagger in a language without fundamental tools, or serve as an improved benchmark for a machine translation model in a language where some data is already available but models are difficult to assess? Tell us your solution in the context of the specific need(s). See Section 3: Purpose and Need for more details.*

Specifications and Deliverables for Proposed Data and Documentation – Please give further information on your proposed effort and state how the proposed quantity/quality of data, collection methods, and other details make the data suitable for use in the operational context of potential use cases. Include the following, *in enough detail that the Technical Advisory Panel is able to assess your proposed work for feasibility, transformative potential, and technical soundness*:

- Quantity, types, and format of data and/or labels, as well as sample frame and size or a plan to ensure representation, if applicable.
- Proposed data collection and labeling techniques and information on interoperability, including consideration of existing or common infrastructure.
- Plans to assess and mitigate error and bias (e.g., gender bias or other biases)
- Metrics to be used to assess desired outcomes of data creation. (i.e. fairness metrics in the dataset, QA/QC against a benchmark, etc.)
- Any anticipated challenges or uncertainties in data collection and proposed countermeasures.

Pathway(s) to Impact and Intended Beneficiaries - Explain how the proposed dataset labeling, creation, augmentation, or maintenance will contribute to achieving impact (for example, using the framework of the Sustainable Development Goals). Describe intended beneficiaries of the application(s) enabled by the dataset and outline previous consultation and/or proposed collaboration with intended beneficiaries.

> *If applicable, describe how the products could motivate multiple and durable paths of research, commercial or non-profit application. For example, associating parallel translation data with images could allow for future research in multiple arenas. Note any practical constraints this pathway to impact may face.*

Accessibility, Data Management, and Licensing – Please describe:

- Plans for data storage, ownership, and discoverability, in line with the Lacuna principles. For example, the proposed data format should be well documented and include information required for the end user to use the data.
- Any anticipated issues related to copyright for source data and collaboration with the copyright holder.
- Plans for licensing to maximize responsible downstream use. Per Lacuna Fund principles, the dataset and any related IP, such as collection methods, datasheets, how to load or read datasets, or other information to ensure usability should be made available under an open

source, by-attribution license (CC-BY 4.0 or similar). If more restrictive licensing is proposed, provide a rationale for this.
- If you intend to use an existing dataset for your project, please indicate that your team has received the necessary permissions from the dataset's owner that the dataset can be released in accordance with Lacuna Fund's IP Policy, or provide justification for another licensing structure.

Risks, Including Ethics and Privacy - Identify potential risks, including but not limited to potential privacy and ethical concerns, and describe steps you will take to mitigate them. Specifically:

- State how you will ensure informed consent if appropriate (this should include notification of potential future use cases for data).
- Describe how you will ensure equity in project labor, including but not limited to fair compensation for labeling and collection and gender parity for field agents.
- Present a plan for anonymization of personally identifiable information (PII) and compliance with privacy laws if applicable. If a national legal framework is not available, the proposal should outline or refer to best practice.
- Discuss potential adverse impacts in the production and use of the dataset and steps to mitigate them.

Sustainability Plan – Describe how the labeled dataset will be maintained, integrated, and/or expanded beyond the initial funding (e.g. through resultant ML applications, by a dedicated community, or a pool of interested parties with a robust governance model for the open dataset).

COVID-19 Considerations – Please state if you will need to adapt your work due to restrictions related to COVID-19. If so, describe the steps you will take to adapt, and in the case of field work, ensure the safety of project staff.

Letters of Support

Letters of support are required from the owner of any existing dataset a proposal intends to use or annotate if information about open licensing for that dataset is not already publicly available. (For example, if a proposal intends to digitize newspaper articles, a letter of support from the newspaper publisher would be required, but if the source dataset is already publicly available and/or licensing terms are clear, a letter of support would not be required). The letter should indicate that the resultant dataset can be released in accordance with Lacuna Fund's IP Policy.

Letters of support from collaborating institutions are optional. If not uploaded to the application portal, they will be requested later from accepted proposals.

Timeline

This section will prompt the applicant to submit a table with a timeline for the completion of major activities and deliverables. The timeline may include, but is not limited to, staff training, data collection, labeling, quality assurance, validation/cleaning, and data publication. Deliverables may include, but are

not limited to, portions of the dataset to demonstrate proof of concept, the full dataset, and accompanying documentation or collection methods to be openly released.

All timelines should include a date by which data will be publicly available with all documentation. Datasets may not be embargoed until publication.

Proposed projects must be completed within 12 months from notification of award (notifications are planned for approximately December 2020/January 2021).

Budget and Budget Narrative

The online application portal will require the applicant to upload a budget formatted in the Lacuna Fund budget template, available on the application page under "Open RFPs" and in the applicant portal. Budgets should be submitted in US Dollars. We are anticipating proposed budgets in the range of $10k – 50k for small to medium-sized projects and as much as $250k for large, complex projects. Note that the Technical Advisory Panel will assess the feasibility and suitability of the budget as well as the linkage between the budget and grant narrative as part of the selection criteria. Budgets may include, but are not limited to, costs for:

- capacity building related to data collection and quality assurance/quality control;
- data collection;
- data labeling;
- QA/QC or verification, including engaging linguists to conduct labeling or ensure quality;
- post-processing of data;
- data publication;
- licensing, including legal advice to ensure copyright and IP compliance;
- costs related to collaborations and consortia;
- community-based methods of data collection or crowd-sourcing efforts, such as label-a-thons.

Funds may **not** be used for the direct payment of any customs, import, or other duties or taxes levied with respect to importation of goods or equipment into any country or jurisdiction. *Indirect costs must be limited to 12% of the proposal budget.*

See the instructions sheet in the budget template for further information on budget guidelines, including information on allowable staff costs.

Budget Narrative: The budget template will also give space to state any budget assumptions or provide additional context.

Annex A: Due Diligence Self-Reporting Form

Submission of this form and all other materials should be completed through the Lacuna Fund application portal, accessible at lacunafund.org/apply.

Meridian Institute must comply with requirements of contributors to the fund as well as legal requirements related to being an organization providing funding based in the United States.

To the extent permitted by law, all information supplied will be held in confidence and not disclosed to any third parties without prior notice and approval.

1. Please state your form of legal incorporation and the relevant law/regulation. Please also provide the month and year when your organisation started operating as this entity.
2. Does your organisation have a Conflict of Interest Policy and are there any instances of Conflict of Interest Meridian Institute should be aware of?
3. Does any Public Official or government entity have any financial, management or controlling interest in your company/organization? If so, provide details and level of interest below.
4. Please list the full names of all Principals for your company/organization. (Note: "Principal" means the executive officers, partners, owners, directors, trustees or others who exercise control over your company/organization).
5. Does your company/organization employ any Public Officials? If yes, please list the name(s) and position(s) of the Public Official(s) below or on a separate sheet of paper.
6. Does any Principal of your organization have a close relative who is a Public Official (close relative means a mother, father, sister, brother, wife or child)? If yes, please list the name, relationship, title, responsibilities, and government department or agency of the Public Official(s) below or on a separate sheet of paper.
7. Has your company/organization, or any subsidiary or affiliate of your company/organization, or Principal of your company/organization ever been convicted in any felony matter or been the subject of a criminal investigation, indictment or similar proceeding? If yes, describe below:
8. During the prior three years, has your company/organization, or any subsidiary or affiliate of your company/organization, been the subject of past or pending litigation, or government investigation? If yes, describe below.
9. Is there additional information about your company/organization or any employee of your company/organization that would assist Meridian Institute in performing its anticorruption due diligence obligations? If so, describe below.
10. The undersigned, being duly authorized to respond to this questionnaire, and new certification as to the matters set forth below, hereby certifies as follows:

    (a) To the best of my knowledge, all information set forth in this anticorruption due diligence form is truthful, correct and complete;

(b) I have read the information concerning the UK Bribery Act 2010 (http://www.justice.gov.uk/downloads/legislation/bribery-act-2010-guidance.pdf), Australian Criminal Code (http://www.oecd.org/daf/anti-bribery/anti-briberyconvention/2027148.pdf), and the U. S. Foreign Corrupt Practices Act http://www.sec.gov/investor/alerts/fcpa.pdf), at the noted websites and I am familiar with the requirements of these anti-corruption statutes;

(c) I have read and fully understand the Meridian Institute Code of Conduct and, as a condition of doing business with Meridian Institute, I agree to abide by such Code of Conduct;

(d) Neither I nor my company have ever paid, approved or otherwise provided anything of value, directly or indirectly, to a Public Official for any improper, corrupt or illegal purpose, nor will we;

(e) Neither I nor my company have ever created a false invoice or otherwise manipulated documentation to disguise making a payment to a Public Official for any purpose, nor will we;

(f) My/our accounting practices will accurately and completely describe any payments made, including any payments made to a Public Official.

NOTE: "Public Official" means any person, whether elected or appointed who holds an executive, legislative, administrative or judicial office or position in any public entity, including any international agency. In addition, "Public Official" includes any person who performs public functions in any branch of the national, state, local or municipal government of any country or territory or who exercises a public function, by employment or under contract, for any public entity, agency or enterprise of such country or territory, including state owned or controlled enterprises. The definition of "Public Official" also includes any official of a political party or any candidate for political office.