Request for Proposals: Natural Language Processing (NLP) 2024

Lacuna Fund: Our Voice in Data

27 June 2024

Table of Contents

1 – Introduction

Overview and Purpose of Lacuna Fund

Lacuna Fund supports the creation, expansion, and maintenance of datasets that enable robust and equitable application of machine learning (ML) 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 the training and/or evaluation of machine learning models.
- 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 fund development and maintenance of equitably labeled datasets most effectively and efficiently.

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. We also see collaboration as a way to enhance the quality and impact of the datasets as well as to promote a culture of cooperation in the field.

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 for proposals is supported by Google.org.

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 as a way to strengthen collaboration and maximize the benefits derived from the use of the datasets, but only the lead applicant will receive funds.

- Have a mission supporting societal good, broadly defined.
- Be headquartered in the country or region where data will be collected. The geographic focus of this call is Africa and Latin America. Institutions based in other countries or regions can apply as partners of the lead institution. As stated above, only the lead applicant will receive funds.
- Have all necessary national or other approvals to conduct the proposed research. The approval process may be conducted in parallel with the 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 proposal – 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 proposals to create, expand, aggregate, and/or unlock datasets for machine learning applications that will enable equitable natural language processing (NLP) outcomes in Africa and Latin America. Lacuna Fund and its partners will perform an initial screen of the proposal for organizational eligibility and feasibility. Following the initial screen, a Technical Advisory Panel of domain experts, data users, and stakeholders will evaluate the 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).

The Technical Advisory Panel for this call will review the submissions and select a set of proposals to be funded. The selections will be based on the degree to which the full proposals meet the following criteria, which are based on the principles that guide Lacuna Fund operations:

- **Quality** – The organization or team proposing the project includes qualified experts in a) the domain of interest; b) machine learning; and c) data management. Collaborations with government agencies and community groups are encouraged. The team provides clear use case(s) for the dataset. The proponent situates the proposed dataset within existing resources (or lack of resources) in the domain and proposes to use effective data collection and labeling techniques and tools to speed the collection, cleaning, and sharing of data.
- **Transformational Impact** – The project makes existing dataset(s) more representative, inclusive, and/or sustainable or creates a new, high-value labeled dataset for an underserved population or problem. A proposal may be considered transformational if it has the potential to address a particularly important or timely problem with equity in ML/AI or have significant reach in terms of number of underserved people or geographies affected.
- **Equity** – The team states the equity issue they are proposing to address and describes how the dataset will fill gaps and make the domain more representative and equitable. There is a compelling theory of change demonstrating how the dataset will create greater access to the benefits of ML/AI for vulnerable and underserved communities.

- **Participatory Approach** – The team is headquartered in the geographic area where the data will be collected to ensure sustained maintenance and usage of the dataset by the local community. In-country partners are involved in strategic elements of the project (beyond data collection). The proposal describes how the team will engage affected stakeholders, seek informed consent for data collection and use, and share project outputs and benefits with data providers and/or the community.
- **Ethics** – The project has a plan and 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. The project's proposed goals and methodology are unbiased and ethical.
- **Sustainability & Communications** – The project has a plan to ensure sustainability and future maintenance of the dataset, e.g., through a baseline model, resultant ML applications, by a dedicated community or a pool of interested parties (for-profit and/or not-for-profit), a robust governance model for the open dataset, and possible machine learning use cases for the dataset. The plan may include who will update and manage the dataset; potential funding sources; proposed engagement strategies for impacted populations and data users; plans to present the dataset(s) at a conference(s); organizing a workshop on dataset sustainability with interested stakeholders; or establish a sustainability committee, as well as measures to keep data open and accessible.
- **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-source licensing, or if this is not possible, a compelling case is made for more restrictive licensing in order to protect privacy or prevent harm. The subgrantee will prioritize releasing the intellectual property under a permissive open-source licensing structure such as Apache 2.0 for any code or other inventions, or CC-BY 4.0 International, or CC BY-SA 4.0 for any other intellectual property (e.g., creative works that are not code, or patentable). The proposed documentation and hosting align with Lacuna Fund's Dataset Hosting and Documentation Guidance.

Timeline

Request for Proposals opens	27 June 2024
Applicant webinar	9 July 2024
Question deadline Please submit questions to secretariat@lacunafund.org	12 July 2024
Mentorship deadline	15 July 2024
Answers posted	29 July 2024
Full proposals due	23 August 2024

Question and Answer Period: All questions related to the RFP should be submitted to secretariat@lacunafund.org with “NLP 2024 RFP Question” in the subject line. Questions submitted by 12 July will be de-identified and answered publicly by 29 July on the Lacuna Fund website in a document posted on the “Apply” page and shared with all the applicants via email.

3 – Purpose and Need

Purpose

The purpose of this call for proposals is to support efforts to develop open and accessible datasets for machine learning applications related to Natural Language Processing (NLP) for low-resource languages and cultures in Africa and Latin America.

The ability to communicate and be understood in one’s own language variety and cultural context is fundamental to digital and societal inclusion. Natural language processing techniques have the potential to enable AI applications that facilitate digital inclusion and improvements in education, finance, healthcare, agriculture, communication, and responses to natural hazards, among others. Many advances in both fundamental and applied NLP have stemmed from openly licensed and publicly available datasets.

However, such datasets are scarce to non-existent for many African and Latin-American languages, excluding these populations from the benefits of NLP. Many current machine learning (ML) models are informed by Anglo-centric and/or translated datasets, lacking culturally relevant nuances and creating biased or unusable models for communities in Africa and Latin America. Where relevant datasets do exist, they are often based on religious or judiciary texts of the past, leading to outdated language and bias. There is a need for openly accessible datasets to facilitate NLP technologies for low-resource languages in Africa and Latin America and support the development of robust and culturally appropriate language datasets that cater to the specific needs of underrepresented communities.

Need

Lacuna Fund seeks proposals from qualified, multidisciplinary teams to develop open and accessible training and evaluation datasets for machine learning applications for NLP in low-resource languages and underrepresented cultures in Africa and Latin America.

Proposals may include, but are not limited to:

- Collecting and/or annotating new data;
- Annotating or releasing existing data;
- Augmenting existing datasets from diverse sources to fill gaps in local ground truth data, decrease bias (such as geographic bias, gender gaps or other types of bias or discrimination), or increase the usability of data and technology related to NLP in low- and middle-income contexts;

- Linking and harmonizing existing datasets (such as across regions, time, linguistic varieties, as well as domain-specific datasets such as historical, health and education data).

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

The TAP sees a need for training and evaluation datasets that will account for the linguistic diversity and cultural nuances in Africa and Latin America. This includes datasets on regional slang, idiomatic expressions, local linguistic varieties or dialects, and culturally relevant data. Such datasets are crucial for developing more inclusive and effective natural language processing tools that can serve the unique needs of culturally diverse linguistic communities.

We seek datasets identified by local experts designed to address locally identified needs. The following are illustrative examples only.

Datasets may include, but are not limited to the following:

- **Labeled and unlabeled datasets for low-resource NLP tasks,** supporting the development of accurate and effective machine learning models. Downstream tasks from labeled datasets might include, but are not limited to: question answering and conversational AI, sentiment analysis datasets, social bias detection, hate speech detection and counterspeech, misinformation and disinformation detection; automatic text summarization or other natural language understanding and generation tasks, or resources to support NLP education in collaboration with communities. Unlabeled datasets include text corpora that can be used to support the training and evaluation of speech models.
- **Speech corpora,** including datasets to enable automatic speech recognition (ASR) that allows illiterate or otherwise underprivileged groups of persons to access information and/or services in low-resource languages.
- **Text-generation tasks datasets,** particularly other than machine translation.
- **Multimodal and other innovative datasets,** such as video or audio captioning, visual question-answering or other image-text interactions.
- **Datasets supporting knowledge-intensive tasks,** such as quality assurance (QA) and Retrieval Augmented Generation (RAG).
- **Datasets related to dialectal variation corpora and code-switched text and speech,** including capturing linguistic variations (regional slang, idiomatic expressions, culturally relevant data) in dialect-rich low-resource languages and in linguistic communities where code-switching is common.
- **Domain-specific creation or augmentation of text and speech datasets,** such as healthcare, place names, agriculture or education, that enable applications with significant social impact. Exploring Generative Data Augmentation frameworks to include domain-specialized vocabulary, semantics, morphology, and syntax.
- **Datasets supporting machine learning for linguistics,** for the preservation and revitalization of marginalized cultures and aspects of underrepresented languages that these cultures consider important for their health, dignity, environment, and well-being. These datasets

may include phonetic, morphological, and syntactic annotations, and automatized tools to perform these tasks if sought by the involved social group

- **Across all datasets: gender-responsiveness and inclusion of key vulnerable groups,** including bias mitigation for those living in humanitarian and conflict settings, as well as those at the intersections of more than one socio-economic group (e.g., disability, gender, age, minorities). Please refer to the 'Risks, including Ethics and Privacy' paragraph on the Proposal narrative section of this document and carefully consider ethics around data collection.

You can review datasets from projects selected in Lacuna Fund's 2020 and 2021 NLP Requests for Proposals to see what work is currently underway.

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 application.

Proposal submissions will only be accepted through the SurveyMonkey Apply application portal available at www.lacunafund.org/apply. Applications can be submitted in English, Spanish, French and Portuguese. A description of application questions is available below for information only. The following sections are required:

- Applicant Information (accessible in the portal)
- Proposal Narrative
- Budget & Timeline

Applicant 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;
- Information about the affiliated institution(s) ethical review processes;
- Information about 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 will not be reviewed.

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

- **Qualifications** - Describe the organization(s) or partnership(s) applying, how they satisfy the eligibility criteria articulated above, and your unique qualifications to undertake the proposed work.
- **Problem Identification and Proposed Solution and Dataset** - Describe the problem or gap in training or evaluation data and the proposed solution. Summarize the dataset(s) you intend to create, augment, aggregate, or maintain. *Please include how your project addresses a gap and complements existing work.*
- **Specifications and Deliverables for Proposed Data and Documentation** – Include the following:
  - Quantity of data that will be included in the dataset.
  - Types and format of data and/or labels, as well as sample frame and size or a plan to ensure representation, if applicable.
  - Metrics to be used to assess desired outcomes of data creation (e.g., fairness metrics in the dataset, QA/QC against a benchmark, etc.)
- **Intended Beneficiaries and Use Cases** - Describe previous consultation and/or proposed collaboration with intended beneficiaries and outline potential current and future machine learning use cases for the proposed datasets. Explain how the dataset respects and reflects the diversity of the communities it represents. State how the proposed quality, collection methods, and other details make the data suitable for use in that particular context.
- **Methodology** – Provide a brief overview of the proposed steps (and key assumptions) for developing and implementing the project. Please include:
  - Proposed data collection and labeling techniques and information on interoperability. Please include consideration of existing or common infrastructure and the latest techniques and tools to speed the collection, cleaning, and sharing of data.
  - Quality control measures, such as the quality all data samples must meet for the dataset. Please include how the team plans to address outliers that may affect the quality of the dataset.
  - A plan to assess and mitigate error and bias (e.g., gender bias or other biases).
  - How you plan to leverage existing resources, including collection methods or technologies, linking pre-existing datasets across the domain, as well as existing resources in other contexts.
  - Permissions in place or steps you will take to secure national or other required approvals. Consider which jurisdictions require approvals and whether the proposed research meets the definition of research in that jurisdiction. If you determine that local, national, or regional approvals are *not* required, please explain why not.
  - Any anticipated challenges or uncertainties in data collection and proposed countermeasures.
- **Transformational Impact** - Explain how the proposed labeling or dataset will contribute to achieving the desired impact. If applicable, describe how the products could motivate use in research or commercial contexts. Note any practical constraints you may face (e.g., internet penetration).
- **Data Management and Licensing** – Please describe:

  - Any anticipated issues related to copyright for source data and collaboration with the copyright holder. Please also address any anticipated issues for copyright and licensing of secondary data.

- Plans for licensing to maximize responsible downstream use. Per Lacuna Fund's Intellectual Property (IP) Policy, 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 CC BY-SA 4.0). If more restrictive licensing is proposed, provide a rationale for this. The budget may include resources for licensing.
  - 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. *Letters of support from existing data holders are optional but encouraged.*
- Plans to include a metadata file and datasheet as documentation for your dataset, according to Lacuna Fund's Dataset Hosting and Documentation Guidance.
- The hosting platform you intend to use. Hosting platforms must assign a digital object identifier (DOI) to the dataset, quantify downloads of the dataset, and collect contact information for dataset downloads. Please see more guidance for suggested hosting platforms in the Dataset Hosting and Documentation Guidance.
- **Risks, Including Ethics and Privacy** - Identify issues and potential risks, including but not limited to potential privacy and ethical concerns, and describe steps you will take to mitigate them. Specifically:
  - A reflective statement on the communities you come from and identities you hold, and how those may impact your work.
  - 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, annotation, provision of data, and other AI services. Some options for prioritizing fair work standards and fair compensation practices include:
- Following and signing the Fair Work Standards pledge for AI work when data annotation and provision are subcontracted to commercial entities.
- Adhering to a community-oriented data provision and annotation approach if such a model exists.
- Presenting a voluntary waiver/survey to project participants to address undue burden.
  - Describe how gender diversity and other demographic considerations are incorporated in the project team, collection of training data, and model development, in order for datasets to accurately represent impacts on different communities/groups.
- 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. Please incorporate privacy considerations at both the individual and community level. Refer to Lacuna Fund's Data Anonymization Guide for suggestions.

- Discuss potential adverse impacts in the production and use of the dataset and steps to mitigate them, including potential human rights risks and the potential for high power consumption in AI technology leading to increased carbon dioxide emissions.
- **Sustainability & Communication Plan**
    - Describe how the dataset will be maintained and/or expanded beyond the initial funding (e.g., through a baseline model, resultant ML applications, by a dedicated community, or a pool of interested parties with a robust governance model for the open dataset) and how a potential use case could sustain the project.
    - Explain how the dataset will follow FAIR data principles (https://www.go-fair.org/resources/faq/what-is-fair/). State the steps you will take to ensure the dataset is **Findable, Accessible, Interoperable, and Reusable.**
    - Outline communications activities to spread awareness of the dataset(s). These could include networking activities with potential data users; presenting the dataset(s) at a conference(s); organizing a workshop on dataset sustainability with interested stakeholders; or establishing a sustainability committee.

Project Timeline & Deliverables

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 open sourced.

All timelines should include a date by which data will be publicly available with all documentation.

Note: Proposed projects must be completed, datasets published, and final reports submitted by October 2026. For planning purposes, you can expect that agreements will be completed and work may begin by April 2025.

Budget

Provide a budget for the completion of the proposed dataset submitted through the SurveyMonkey Apply portal. This should be formatted in the Lacuna Fund budget template, available in the applicant portal.

The total pool available is approximately $1 million USD. We would like to fund projects in each of the target regions (Africa, Latin America) and anticipate supporting 6-8 smaller projects with budgets up to $100k USD and 2-3 larger, more complex projects with budgets ranging from $100-250k USD. 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 (including fair compensation for data provision);
- data labeling (including fair compensation for data labeling);

- QA/QC or verification;
- post-processing of data;
- data publication;
- licensing;
- open access publication of results;
- time to prepare a data statement for the dataset;
- crowd-sourcing efforts, such as label-a-thons;
- data storage;
- computing power;
- workshop
- communications activities, including conference attendance for up to two events to present the datasets

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 rates are strictly limited to 14.5% of direct research costs.**

In-kind cloud storage and computing power may be available from Lacuna Fund partners. If you would like to utilize this resource, please include it in your budget. Selected teams will receive instructions for how to apply when they receive their award.

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

Thank you for your interest in Lacuna Fund and your efforts to support greater equity and accessibility for machine learning applications to support natural language processing. We look forward to reviewing your submission!

NOTE: Mentorship & Matchmaking Opportunity

Lacuna Fund is pleased to partner with Masakhane Research Foundation (MRF) to offer mentorship and matchmaking opportunities for NLP applicants. MRF is a grassroots organization whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans. MRF's goal is for Africans to shape and own these technological advances towards human dignity, well-being and equity, through inclusive community building, open participatory research and multidisciplinarity.

For this Lacuna Fund call for proposals, MRF will offer applicants from Africa and Latin America a special opportunity to join the MRF community and match those who are interested with a mentor who will review your draft proposal and discuss avenues for strengthening it. Interested parties may apply for a session with a mentor by filling out this Google Form with a brief description (250-words abstract) of the dataset proposal. Mentees can request different forms of assistance such as ***"discussing gaps in low-resource NLP," "writing a research proposal,"*** and ***"preparing budgets."*** Providing a detailed description will support matching applicants to mentors that are aligned in your interest area. Matching of mentors to mentees will be administered by our Partner Matchmaking and Mentorship Associates who will organize several workshops to bring together applicants who are working in related fields (e.g. linguists and NLP researchers) and coordinate and organize communication sessions between relevant groups. Through this process, we will also provide match-making opportunities for applicants to collaborate with each other to form project teams and submit proposals together.

Interested applicants are encouraged to apply for a mentorship session at least 6 weeks before the proposal due date, by **15th July 2024**. We will do our best to identify a mentor for everyone who requests one, but we cannot guarantee that mentors will be available for all. Mentors will be assigned on a first come, first-served basis. All applicants are expected to read and abide by the mentorship programme's *code of ethics and conduct*.