

Inclusive Datasets for Machine Learning

**Natural Language Processing 2024
Applicant Webinar**

July 9th

Lacuna Fund Hub



**Managing the 2024
NLP call for
proposals**

Based in Chile



**ALVARO SOTO
CENIA**



**CRISTINA FLORES
CENIA**

Speakers



KATRINA GEHMAN
MERIDIAN INSTITUTE,
LACUNA FUND SECRETARIAT



DR. ALBERT KAHIRA,
DATA QUALITY ADVISOR,
DATAWISE AFRICA



MENTORSHIP & MATCHMAKING,
MASAKHANE

Technical Advisory Panel (TAP) Members



**AUDREY JULIA WALEGHWA
MBOGHO
USIU-AFRICA**



**FELIPE DANIEL HASLER
SANDOVAL, PHD
UNIVERSIDAD DE CHILE**

Meeting Objectives



NLP 2024

- Provide potential applicants with an understanding of Lacuna Fund and proposal requirements.
- Provide time for applicant questions about the Request for Proposals (RFP).

Agenda

- 00:00** **Welcome, Introduction, Agenda Review**
- 00:10** **Presentation: Lacuna Fund Overview and Requirements for RFPs**
- 00:40** **Presentation: Data Quality and Hosting Considerations**
- 00:50** **Presentation: Mentorship/Matchmaking Opportunity**
- 01:00** **Question and Answer**
- 01:25** **Next Steps**
- 01:30** **Adjourn**



A lacuna is a gap, or a missing portion of a manuscript



Lacuna Fund supports the creation, expansion, and maintenance of equitable training and evaluation datasets that enable machine learning tools to better tackle urgent problems in low- and middle-income contexts globally.

Principles that guide Lacuna Fund:

- Accessibility
- Equity
- Ethics
- Participatory Approach
- Quality
- Transformational Impact



Funders



Lacuna Fund Governance Structure

Steering Committee

Strategic guidance and overall governance of the Fund

Funder Collaborative

Provides input on key priorities to the Steering Committee and provides a forum for deeper funder collaboration

Technical Advisory Panels

Scope funding processes and select proposals, provide input on domain specific strategies and needs

Secretariat

Manages grants and reporting, provides facilitation, communications, and operational support



Domains for Datasets

Lacuna Fund provides resources for dataset creation, expansion, or maintenance.

Current calls for proposals in these domains:

- **Natural Language Processing (NLP)**
- **Antimicrobial Resistance (AMR)**

Prior calls for proposals in these domains:

- Agriculture
- Language
- Health
- Climate

Future calls for proposals:

- Other areas where the Steering Committee identifies a need



Funder for NLP 2024

The 2024 NLP call for proposals is made possible with generous support from [Google.org](https://www.google.org).



Google.org



Requirements for RFP

NLP 2024



NLP 2024 RFP - Purpose

Stated 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 in Latin America and 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 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 unusable models for communities in Latin America and Africa. **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 in African and Latin American low-resource languages** and support the development of robust and comprehensive language datasets that cater to the specific needs of underrepresented communities.

NLP RFP – 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. Datasets may include, but are not limited to the following:

- Labeled and unlabeled datasets for low-resource NLP tasks
- Speech corpora
- Text-generation task datasets
- Multimodal and other innovative datasets
- Datasets supporting knowledge-intensive tasks
- Datasets related to dialectal variation corpora and code-switched text and speech
- Domain-specific creation or augmentation of text and speech datasets
- Datasets supporting machine learning for linguistics
- Across all datasets: gender-responsiveness and inclusion of key vulnerable groups.



What We Are Looking For

- 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).





Proposal Information and Eligibility Requirements

Proposal Information

- Applicant Information
- Proposal Narrative
- Project Timeline & Deliverables
- Budget
- NOTE: Mentorship & Matchmaking Opportunity



Eligibility Requirements



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.

Eligibility Requirements



To be eligible for funding, organizations must:

- 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.

Key Dates

27 June 2024: Request for Proposals Open

9 July 2024: Applicant Webinar

12 July 2024: Questions deadline

Please submit questions to secretariat@lacunafund.org

15 July 2024: Mentorship deadline

29 July 2024: Answers posted

23 August 2024: Full proposals due

Spring 2025: Grants awarded

***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.**

Elements of Strong Proposals



Elements of Strong Proposals

- **Multi-disciplinary team** - experience in data science, Natural Language Processing (NLP)
- **Use cases and community engagement**
- Clear **problem statement** and how the proposed dataset or aggregation will help solve the problem
- Specificity about dataset size – should be of **sufficient size and quality** to be useful in future applications
- **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.



Elements of Strong Proposals

- **Equity** considerations (gender, socio-economic status, ethnicity, etc.)
- Consideration and plan for potential **privacy and ethics** issues
- Plan for **data management and licensing**
- Working across areas where possible and relevant (i.e. regions, time, linguistic varieties)
- **Budget** is appropriate for the size of the dataset produced



Community Engagement

- Describe previous consultation and/or proposed collaboration with intended beneficiaries.
 - When and where you have met/ will meet with partners
 - How partners will be involved in:
 - Identifying data needs
 - Data collection, labeling
 - Data governance
 - Dataset use
 - How partners will benefit from the new/expanded dataset



Privacy and Ethics

- Explain how your team will address:
 - 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.
- Describe the process you will use to screen for potential ethical issues (e.g., an institutional review board, etc.).



Sustainability & Communications 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.
- 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.

Data Standards and Sharing



- Findable - easy to find on a public, widely used platform
- Accessible - open format (CCBY 4.0 or CC BY SA 4.0)
- Interoperable - data format
- Reusable - metadata

- Sustainable - plan for maintenance
- Shared – engaged community of users & plan for sharing dataset once complete

<https://www.go-fair.org/fair-principles/>

Lacuna Fund Intellectual Property Policy: [IP-Policy_LacunaFund.pdf](#)

Budget and Allowed Costs

- 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.

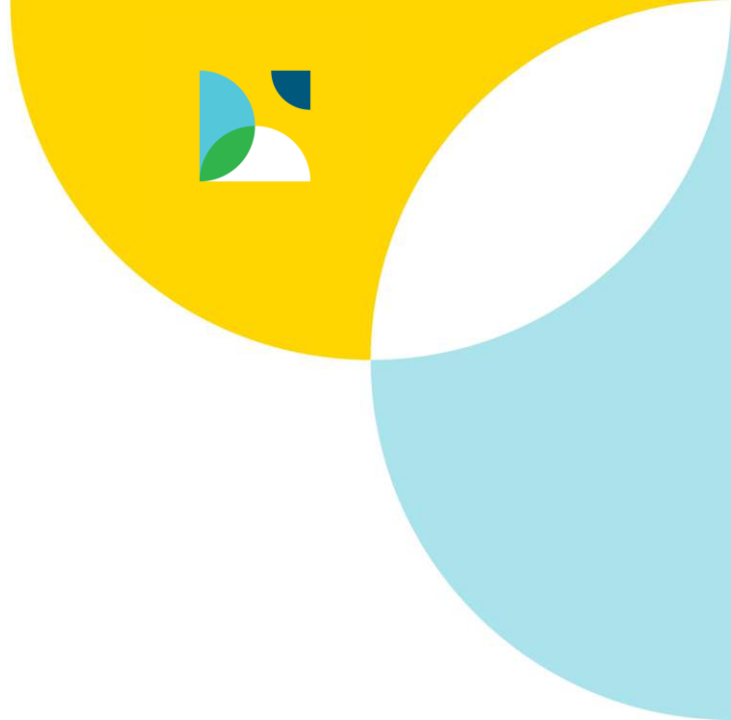
Budget and Allowed Costs



Budgets may include, but are not limited to, costs for:

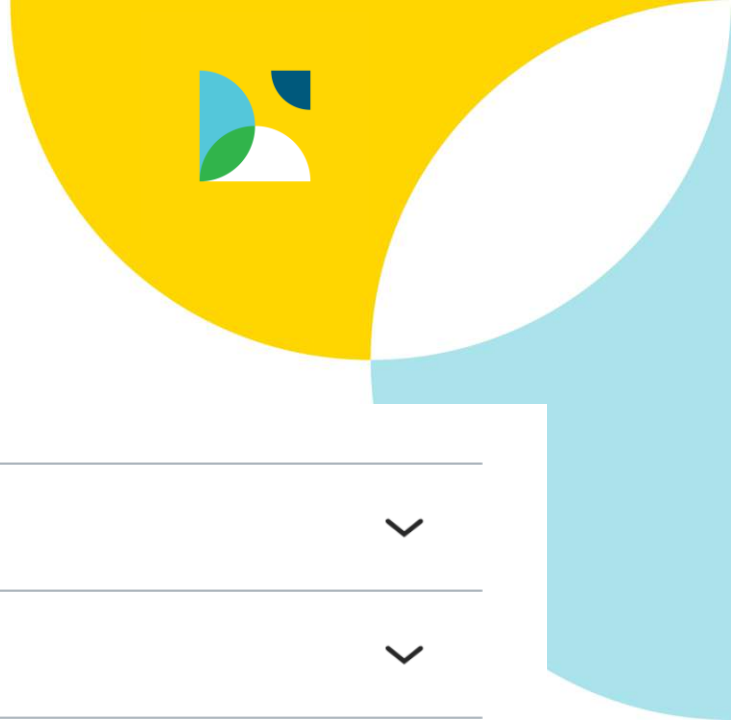
- Capacity building related to data collection and quality assurance/quality control;
- Data collection (including fair compensation for data provision);
- Data labeling (including fair compensation for data labeling);
- QA/QC or verification;
- Post-processing of data;
- Data publication;
- Baseline Model

- 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



Data Quality Considerations

Hosting and Accessibility –Lacuna Principles



| Accessibility



| Equity



| Ethics



| Participatory Approach



| Quality



| Transformational Impact



Hosting – Lacuna Guidelines

- Assigns a digital object identifier (DOI) for datasets or allows one to be attached as part of the metadata.
- Is indexed by major search engine(s) (e.g., Google Dataset Search or similar tools).
- Is reliable and persistent

Great to have:

- Quantifies the number of landing page views and downloads for the dataset.
- Collects contact information for dataset downloads in a way that maximizes conversion.



Dataset Hosting

The proposed documentation and hosting aligns with Lacuna Fund's [Dataset Hosting and Documentation Guidance](#).

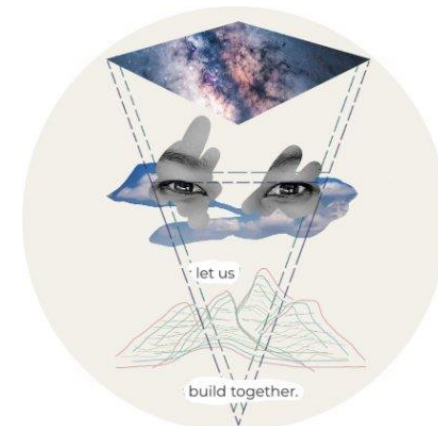
Lacuna Fund asks grantees to include the following documentation when datasets are submitted:

- Metadata file
- Datasheet
- Digital object identifier (DOI)



Mentorship Opportunity: The Masakhane Mentorship Programme Opportunity:

Masakhane Research Foundation



The Masakhane Mentorship Programme



- **Mission:** 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 multidisciplinary.
- 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.

The Masakhane Mentorship Programme

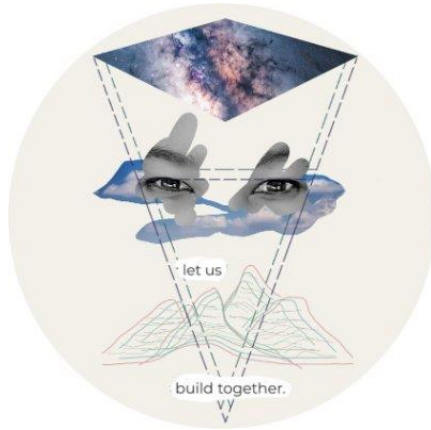


- Apply on the **Google Form** below <https://forms.gle/8u5GobjKXYo7gujt5>
- Provide a brief description (250-words abstract) of the dataset proposal
- Mentees can request for different forms of assistance such as
 - “discussing gaps in low-resource NLP”,
 - “writing a research proposal”, and
 - “preparing budgets”.

The Masakhane Mentorship Programme



- Interested applicants are encouraged to apply for a mentorship session at least 6 weeks before the proposal due date, by **15th July 2024**
- 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](#).



Please find more information
at: <https://lacunafund.org/apply/>

Questions?

Email:

masakhane_leadership@googlegroup.com

Lacuna Fund NLP call: Request for mentorship from Masakhane Research Foundation

Lacuna Fund is pleased to partner with Masakhane Research Foundation (MRF) to offer mentorship opportunities for applicants.

For this Lacuna Fund call, 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.

Pour le **Français** : veuillez visiter <https://forms.gle/PhChp4vcX4ctx1my9>

Para **Español**: visite <https://forms.gle/PvToC1PpsuGTdtQk8>

Para **Português**: visite <https://forms.gle/shBcUtVS98uWaBBh9>

davlanade@gmail.com [Switch account](#)

Not shared

* Indicates required question

Proposal title *

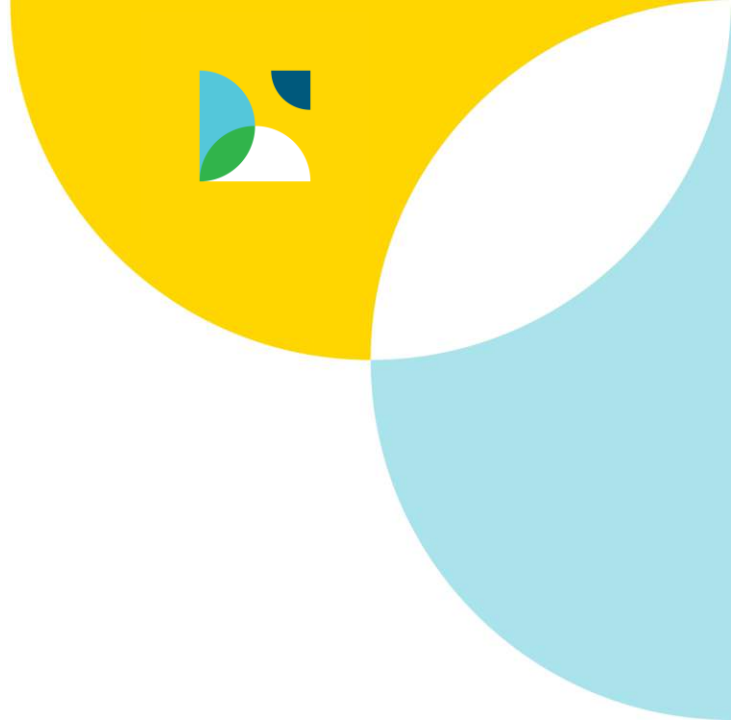
Your answer

Proposal abstract (250-words) *

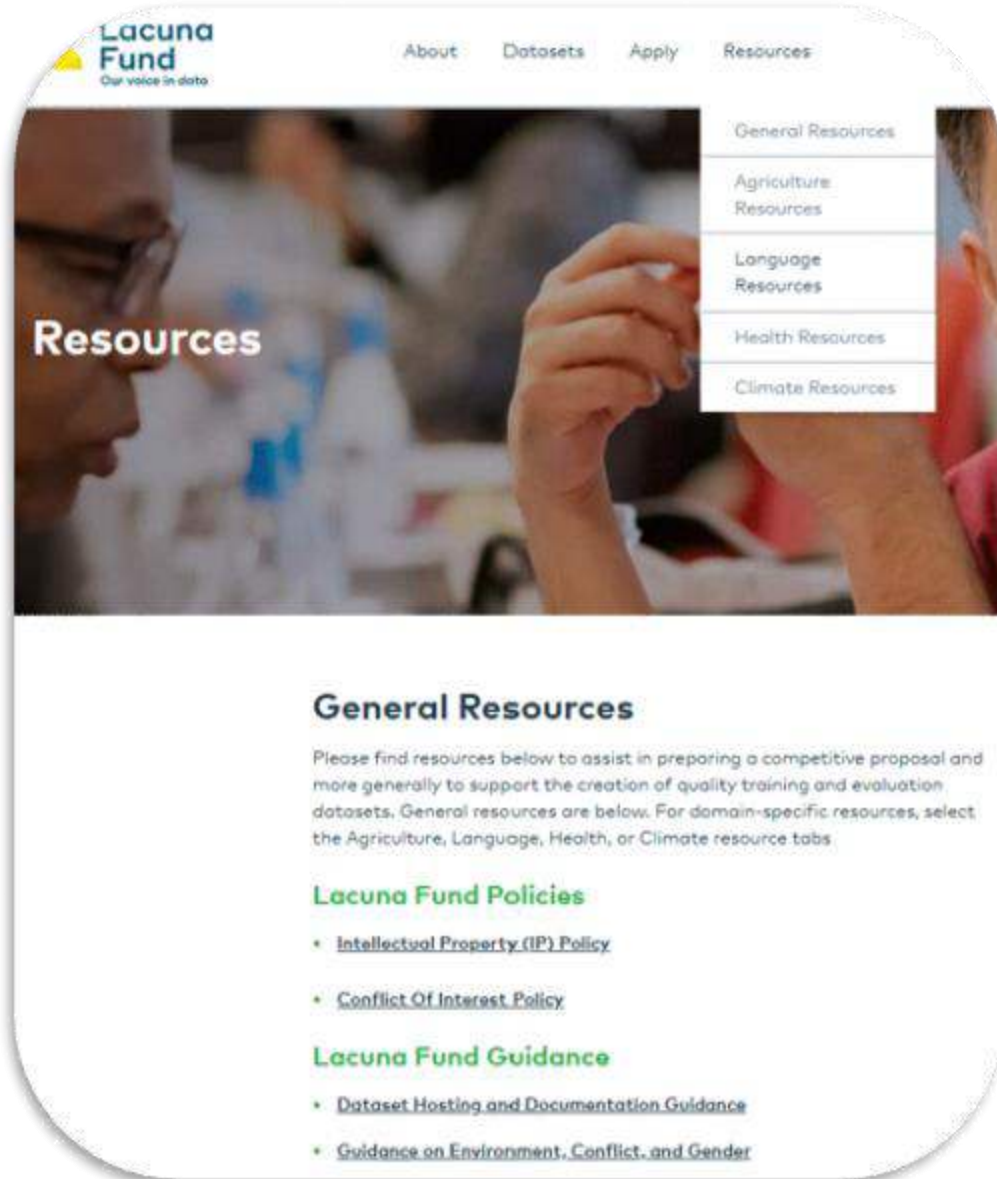
Your answer

Principal investigator (PI) *

Resources



Resources



Lacuna Fund
Our voice in data

About Datasets Apply Resources

Resources

- General Resources
- Agriculture Resources
- Language Resources
- Health Resources
- Climate Resources

Resources

General Resources

Please find resources below to assist in preparing a competitive proposal and more generally to support the creation of quality training and evaluation datasets. General resources are below. For domain-specific resources, select the Agriculture, Language, Health, or Climate resource tabs.

Lacuna Fund Policies

- [Intellectual Property \(IP\) Policy](#)
- [Conflict Of Interest Policy](#)

Lacuna Fund Guidance

- [Dataset Hosting and Documentation Guidance](#)
- [Guidance on Environment, Conflict, and Gender](#)



Lacuna Fund
Our voice in data

About Datasets Apply Resources

Language Resources

Resources for Proposals in NLP

This document of [2024 NLP Resources](#) (also listed below) represents a collection of resources from the Technical Advisory Panel (TAP) as an addition to those referenced in the RFP document. These are intended to provide assistance in obtaining relevant background information, preparing a competitive proposal, and completing quality work.

These resources are not intended to be exhaustive nor authoritative. This document does not represent an endorsement of work by the Lacuna Fund Secretariat, the TAP, or individual members.

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

Proposal Submission

Proposal submissions will only be accepted through the SurveyMonkey Apply application portal available at www.lacunafund.org/apply.

Applications can be submitted in **English**, **French**, **Portuguese**, and **Spanish**.

Note: Select your desired language by using the dropdown tab in the application portal. For those submitting an application in Portuguese, you may apply using either the English, Spanish, or French portals. We do not have a Portuguese portal option available at this time. However, proposals submitted Portuguese in any portal are accepted and will be reviewed

SurveyMonkey Apply (SMA) Portal



Meridian Institute

Lacuna Fund: Natural Language Processing 2024

Lacuna Fund is the world's first collaborative effort to provide data scientists, researchers, and social entrepreneurs in low- and middle-income contexts globally with the resources they need to produce labeled datasets that address urgent problems in their communities. Please visit lacunafund.org for more information about the Fund.

Lacuna Fund seeks proposals from organizations to develop open and accessible training and evaluation datasets for machine learning applications in natural language processing (NLP) in low- and middle-income countries around the world. The RFP closes on 23 August 2024 at 11:59 PM, US Mountain Daylight Time. (GMT-7 hours)

Please find the [full RFP available on our website](#).

APPLY

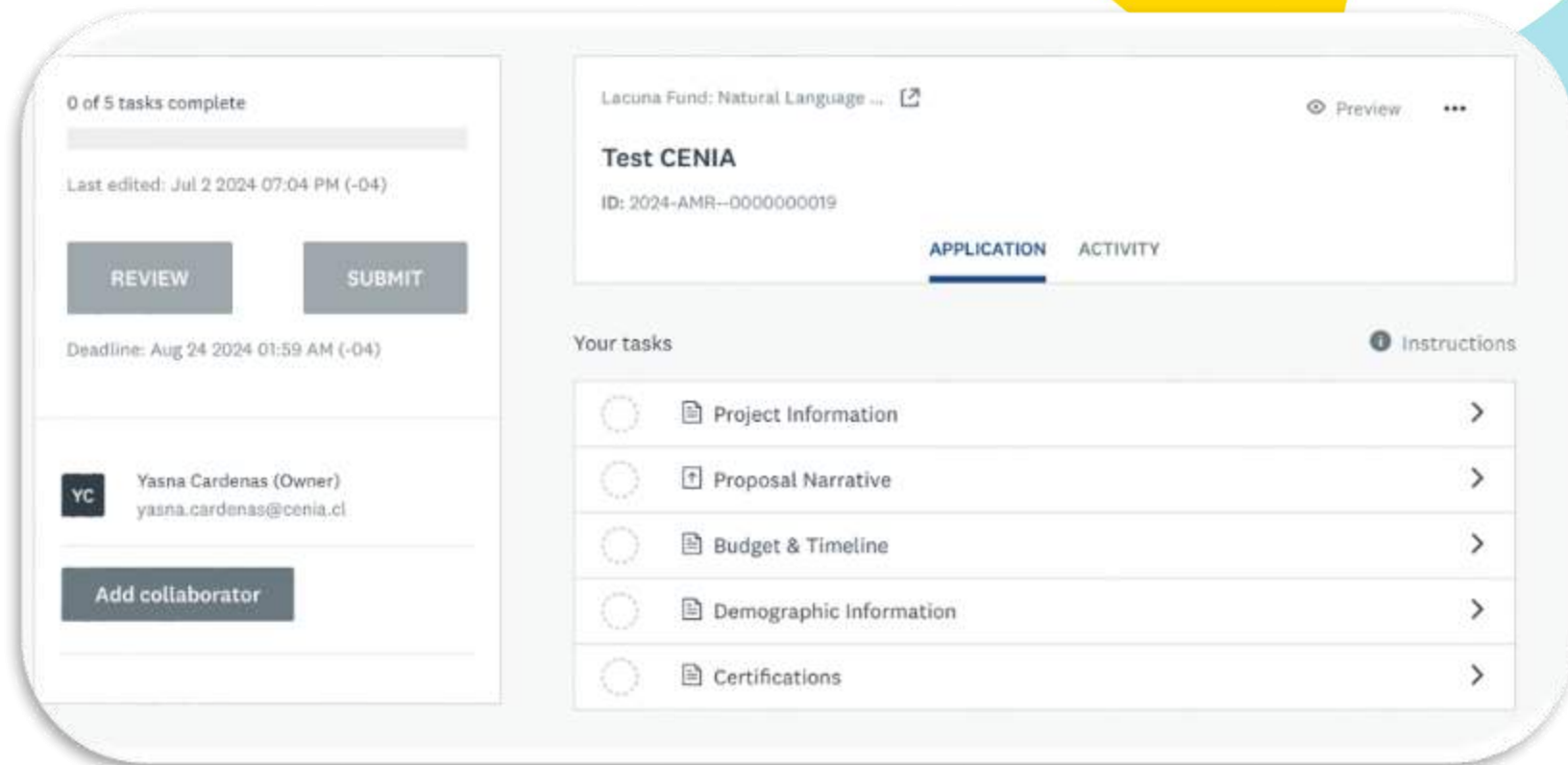
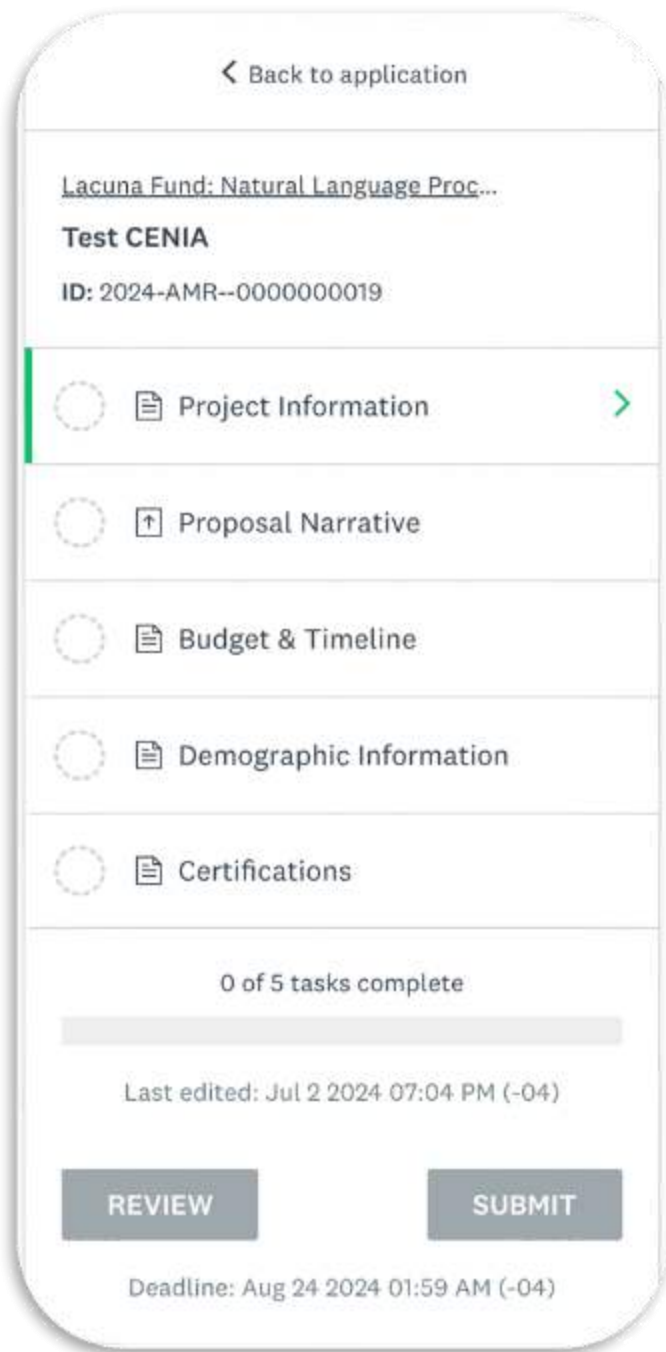
Opens

Jun 27 2024 12:00 AM (MDT)

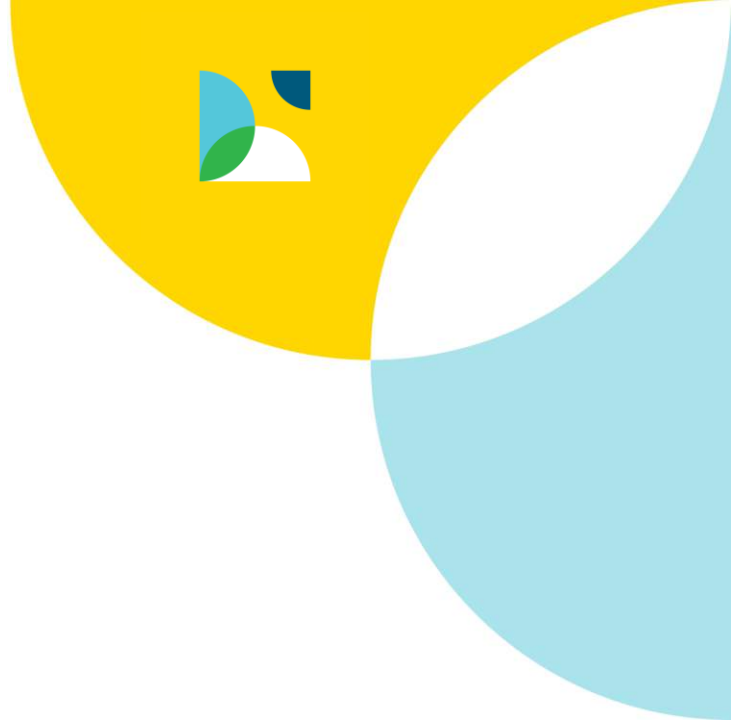
Deadline

Aug 23 2024 11:59 PM (MDT)

Application



Questions?



Next Steps

- **Submit additional questions to secretariat@lacunafund.org by 12 July 2024.**
- **Answers posted** publicly on Lacuna Fund Apply page on **29 July 2024**
- **Proposals due on 23 August 2024.**