Lacuna Fund
Our voice on data

Expressions of Interest: Labeled Agricultural Datasets for Machine Learning Solutions in Sub-Saharan Africa

Lacuna Fund: Our Voice on Data

6th Aug 2021

Table of Contents

1 – Introduction

Purpose and Goals of Lacuna Fund

Lacuna Fund supports the creation, expansion, and maintenance of equitably labeled datasets that enable the robust 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.

Principles of Lacuna Fund

The following principles guide the operations of the Lacuna Fund.

- **Accessibility** – The Fund is committed to ensuring that labeled 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 –** The Fund aims to make AI more equitable by supporting the creation, expansion, and maintenance of 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 the systems or technologies that create harm.
- **Ethics –** The Fund will fund data collection in a manner consistent with ethical labor standards and requires sub-grantees 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 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** – 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.**

This call for proposals is supported by Canada's International Development Research Centre, GIZ on behalf of Germany's Federal Ministry for Economic Cooperation and Development, Google.org, and The Rockefeller Foundation.

2- Overview

Those that submitted in 2020 are welcome to submit again. Applicants proposing to build on one of the datasets funded in 2020 should explain what value will be added by new work. You may find information about datasets funded in 2020 on the Lacuna Fund website.

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 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 welcome but only the lead applicant will receive funds.
- Have a mission supporting societal good, broadly defined.
- Be an organization headquartered in Africa or have a substantial partnership with an organization(s) headquartered in Africa. Regional presence in Africa is not a sufficient condition, but a substantial partnership with those headquartered in Africa is.
- 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 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.. agriculture) to allow high-quality AI/ML analytics to be performed by multiple entities.

Selection Process and Evaluation Criteria

Lacuna Fund seeks to hear from organizations that are interested in responding to a call for proposals to unlock, create, aggregate, and/or improve labeled datasets that will be essential to train and improve machine learning (ML) models needed for sustainable transformation of the agri-food systems in sub-Saharan Africa. The Fund and its partners will perform an initial screen of Expressions of Interest (EOI) for organizational eligibility and feasibility of the original concept presented in the EOI. Following the initial screen, a Technical Advisory Panel of domain experts, data users, and stakeholders will evaluate the EOIs based on the selection criteria outlined below. Lacuna Fund will then invite, through a Request for Proposals (RFP), a small number of organizations to provide a full proposal based on the EOI submission. 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 shortlist of organizations that will then be invited to provide full proposals for funding. Selections for the shortlist of invited applicants will be based on the degree to which they meet the following criteria:

- **Accessibility –** The dataset will be made: a) Widely accessible under open access licensing pursuant to Lacuna Fund's IP Policy. If this is not possible, a compelling case is made for more restrictive licensing to protect privacy or prevent harm along with a mechanism for providing access under proposed licensing. b) Easily discoverable by hosting them on a permanent repository such as the Radiant's MLHub or similar platforms.
- **Efficiency –** The proponent has considered current practice and existing datasets and proposes to use effective data collection, labeling, and standard formatting techniques and tools to speed the collection, cleaning, and sharing of data.
- **Equity –** There is a compelling theory of change demonstrating how the dataset will improve machine-learning solutions and be applied to help vulnerable and underserved communities.
- **Ethics –** The project can 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.
- **Feasibility –** The project is feasible in relation to the budget and scope of work proposed. *Feasibility will be considered at the EOI stage but will be more comprehensively evaluated at the full proposal stage for those selected to proceed to that round.*
- **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.

- **Quality –** The organization or team proposing the project includes qualified experts in a) the agricultural area of interest; b) machine learning; and c) data management.
- **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.
- **Transformational Impact –** There is a clear demonstration of how the proposed work is scalable and/or transformational. The work could fall in either of the two categories:
    - A. Creating new datasets - It creates a new, high-value labeled dataset for an underserved population or problem related to the Sustainable Development Goals (SDGs);
    - B. Building on existing datasets - It either a) labels or collects validation data for existing labeled datasets or upcoming surveys and in so doing, unlocks additional value in the existing dataset; b) makes an existing dataset more representative and inclusive of low- and middle-income contexts; or c) makes a widely used and equitable dataset more sustainable.

Timeline

EOI Call Posted Publicly on Lacuna Fund Website	6th Aug 2021
Question and Answer Deadline Please submit questions to secretariat@lacunafund.org	16th Aug 2021
Answers Posted	23rd Aug 2021
Expressions of Interest Due	3rd Sep 2021

Question and Answer Period: All questions related to the EOI should be submitted to secretariat@lacunafund.org with "Agriculture EOI 2021 Question" in the subject line. Questions submitted by 16th August 2021 will be de-identified and answered publicly by 23rd August 2021 on the Lacuna Fund website in a document posted on the "Apply" page.

3 - Purpose and Need

This call for expression of interest and the final request for proposals will fund labeled agricultural datasets for machine learning in sub-Saharan Africa, whether referenced to earth observation (EO) data or related to other aspects of the crop and animal agricultural system. Projects focused on building ML models or other applications will not be considered.

Purpose

The purpose of this call for EOI is to identify viable projects to submit full proposals to develop open and accessible training and evaluation datasets for machine learning (ML) applications that address the challenges in the agricultural sector in sub-Saharan Africa. The sector is largely defined by resource-

poor, rain-dependent, small-scale production systems, and data is scattered all over. Novel and highly efficient ML solutions are increasingly essential towards accelerating the much-needed agricultural transformation. The value of ML in agriculture is its ability to collect and/or process huge datasets beyond the scope of human capability and then reliably convert analyzed data into decision-support information for players along the agricultural value chain. It ultimately leads to better outcomes: lower costs, increased profits, increased food security, reduced poverty, and increased wellbeing. More timely, accurate and complete datasets with standard labeling and reflexive documentation of their creation processes are needed to train and evaluate the ML models and enhance their use potentials. Equitable implementation of highly representative ML models has the potential to improve the efficiency of different activities in the value chain by ensuring more informed decision-making among respective actors.

In the agricultural AI for social good domain, recent advances in the analysis of remote sensing data have enabled improved accuracy of a variety of tasks, from the prediction of crop types and yields to field boundary estimation worldwide.[1] Both startups and large development programs have built on this data to develop ML applications that allow for both more personalized services and better data for decision-making.

However, a lack of ground truth labels, as well as a lack of datasets to address unique challenges in mapping smallholder farms, hinders further progress towards building beneficial ML applications that are effective for under-served populations worldwide.[2] Alongside a need for greater open access, quality data, an African ecosystem capable of building on the data in real-world applications must be strengthened to ensure equitable benefit from any funding.

An explosion of EO providers and services, nascent standards, such as the Spatio-Temporal Asset Catalog (STAC)[3] and a preliminary version[4] of best practices for ground reference data collection and cataloging are creating greater coherency for ML-ready labeled data across the AI for agriculture landscape.

Need

Lacuna Fund seeks Expressions of Interest (EOIs) from qualified organizations to develop open and accessible training and evaluation datasets for ML applications that address the challenges in agricultural value chain in sub-Saharan Africa. The agricultural value chain could be split into four major phases namely i) pre-production, ii) production, iii) processing, and iv) marketing and distribution

[1] Lei Ma et al., "Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review," *ISPRS Journal of Photogrammetry and Remote Sensing* 152 (June 1, 2019): 166–77, https://doi.org/10.1016/j.isprsjprs.2019.04.015.

[2] Jadunandan Dash, "Satellites and Crop Interventions," *Nature Sustainability* 2, no. 10 (October 2019): 903–4, https://doi.org/10.1038/s41893-019-0402-3.

[3] https://github.com/radiantearth/stac-spec

https://github.com/radiantearth/ground-referencing-guide

phases. Extension services and inputs supply and trade functions cut across and are critical for the performance of different actors throughout the value chain. Different sets of datasets are needed for efficient decision-making among actors within and across the phases of the value chain. This demonstrates the complexity of activities in the sector and the need for highly integrated information for effective decisions.

Traditionally such information would be delivered to farmers through the extension agents. However, there are insufficient numbers of extension agents in the region to meet the need. Currently, one extension officer is to one thousand five hundred farmers (1:1500) as compared to the FAO recommended ratio of 1:400. Knowledge and information access and utilization by the expected recipients have since become impossible- a gap that can be filled by the use of ML. Unlike extension officers who could advise and provide end-to-end knowledge and information and which was used to support decision-making, ML solutions have not covered all the datasets across the value chain in a manner that the dataset could be used to provide end to end knowledge and information enough to aid decision making.

Some of the specific AI/ML application areas that require representative and complete labeled datasets for effective use in the region include:

Soil, Water and Crop Management

a. Soil properties prediction;
b. Water and irrigation management;
c. Yield estimation;
d. Crop type classification;
e. Crop protection;
f. Weed management:

Livestock Management

g. Breeding and livestock breeds dataset
h. Livestock management;
i. Feed optimization and forage availability prediction;
j. Livestock diseases and management
k. Prediction and early warning on the anticipated conflict between pastoralists and crop farmers;
l. Management of pastoral migratory patterns among others.

Cross-cutting categories

m. Prediction and early warning on the anticipated conflict between pastoralists and crop farmers;
n. Input supply and demand management;
o. Produce supply chain management;
p. Produce aggregation and demand management;
q. Input and produce marketing and distribution management;
r. Produce processing;
s. Extension service provision;

Some of the most valued crops and livestock, produced and kept by smallholder farmers, which contribute to food security include maize, wheat, rice, cassava, pulses, kales, indigenous crops, chicken, cattle, goat, pigs, sheep, fish and honey bees. Lacuna Fund's call for proposals is intentionally open to funding ML datasets covering and applicable beyond areas and commodities mentioned above.

The Fund would consider projects, which look into the mentioned needs and which incorporate as many as possible of the dynamics and challenges faced by the smallholder farmers in sub-Saharan Africa. However, the TAP emphasizes the need for clarity on the scope of the proposed datasets - including geography, commodity, source(s), and use potentials - and how these projects align with the Lacuna Fund principles as illustrated above (page 2).

Project Timing – The TAP recognizes that the seasonality of agriculture affects when data can be collected. The Fund will consider projects that have both data labeling and/or collection aspects, Applicants are encouraged to embrace reflexive documentation of the datasets and demonstrate their regard for dataset transparency and accountability to facilitate effective communication between dataset creators and dataset users. Applicants may consider the datasheet[5, 6] questions to enhance the quality of their submissions and data creation activities.

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 2 pages not including references, with 2.5 cm margins, single-spaced and a minimum of 11-point font. Submissions exceeding 4 pages (including references and appendices) may not be considered.

The following information is required:

- **Applicant Information**
  - Name of Participating Organizations
  - Descriptions and Qualifications of Participating Organizations – Provide a 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 agricultural datasets.
- **Problem and Proposed Solution**

[5] https://www.microsoft.com/en-us/research/project/datasheets-for-datasets/
[6] Gebru, T. et al. (2020). Datasheets for Datasets. Cornell University. Available at https://arxiv.org/pdf/1803.09010

- Problem Identification and Proposed Solutions – Describe the agricultural sector players' inequality and disparity, which is affected most by this problem, and the proposed solution (dataset labeling, aggregation, creation, augmentation, or maintenance).
- General Methodology – Provide a brief overview of the proposed steps (and key assumptions) for developing and implementing the solution set. Explain permissions in place or steps you will take to secure national or other required approvals.
- Expected Outcomes and Benefits Following Project Implementation - Explain how the proposed project will contribute to achieving the desired impact. If applicable, describe how the products could motivate multiple and durable paths of research or commercial application.

- **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 the total number of months required for completion.
  - **Note**: Proposed projects must be completed, datasets published, and final reports submitted no later than January 31, 2024. For planning purposes, you can expect that agreements will be completed and work may begin by March 2022.
  - Budget – Provide a broad overview of the expected budget for the completion of the steps included in the General Methodology above. Budgets should be submitted in US Dollars. The total pool available is approximately USD 800,000. We plan to fund projects with a maximum budget of USD 200,000. For information for planning purposes, the indirect rate is 12%.

Those who are invited to submit full proposals will be asked to submit a detailed budget and timeline; describe specifications for data collection, storage, and management; and provide a basic risk mitigation plan in the context of COVID-19. For earth observation data, metadata should be listed in the STAC standard. Recipients of funding will be asked to prepare a datasheet for their dataset. 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, rationale will be requested.

Thank you for your interest in Lacuna Fund and your efforts to leverage ML applications in the agriculture sector. We look forward to reviewing your submission.