Conjuntos de dados inclusivos para aprendizado de máquina

Processamento de linguagem natural 2024
Webinar para candidatos

9 de julho

Lacuna Fund

Centro de Fundos Lacuna

CENIA

CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL

ALVARO SOTO
CENIA

CRISTINA FLORES
CENIA

Gerenciando a chamada de propostas do PNL 2024

Com sede no Chile

Alto-falantes

KATRINA GEHMAN

INSTITUTO MERIDIAN, SECRETARIA DO FUNDO LACUNA

DR. ALBERT KAHIRA, CONSULTOR DE QUALIDADE DE DADOS, DATAWISE ÁFRICA

MENTORIA E MATCHMAKING, MASAKHANE

Membros do Painel Técnico Consultivo (TAP)

AUDREY JULIA WALEGHWA MBOGHO

USIU-AFRICA

FELIPE DANIEL HASLER SANDOVAL, PHD

UNIVERSIDADE DO CHILE

Objetivos da reunião

NLP 2024

- Fornecer aos possíveis candidatos um entendimento do Fundo Lacuna e dos requisitos da proposta.

- Dê tempo para que os candidatos façam perguntas sobre a Solicitação de Propostas (RFP).

Agenda

00:00	**Boas-vindas, apresentação,** revisão da **agenda**
00:10	**Apresentação: Visão geral do Fundo Lacuna e requisitos para RFPs**
00:40	**Apresentação: Considerações sobre qualidade de dados e hospedagem**
00:50	**Apresentação: Oportunidade de Mentoria/Matchmaking**
01:00	**Perguntas e respostas**
01:25	**Próximas etapas**
01:30	**Encerrar**

Uma lacuna é uma lacuna ou uma parte ausente de um manuscrito

O Lacuna Fund apoia a criação, a expansão e a manutenção de conjuntos de dados de treinamento e avaliação equitativos que permitem que as ferramentas de aprendizado de máquina resolvam melhor os problemas urgentes em contextos de baixa e média renda em todo o mundo.

Princípios que orientam o Lacuna Fund:

- Acessibilidade
- Patrimônio líquido
- Ética
- Abordagem participativa
- Qualidade
- Impacto transformacional

Financiadores

The ROCKEFELLER FOUNDATION

Google.org

IDRC · CRDI
International Development Research Centre
Centre de recherches pour le développement international

Canada

german cooperation
DEUTSCHE ZUSAMMENARBEIT

Federal Ministry for Economic Cooperation and Development

FAIR FORWARD
Artificial Intelligence for all.

GORDON AND BETTY MOORE FOUNDATION

Patrick J McGovern FOUNDATION

Robert Wood Johnson Foundation

wellcome

Fundo Lacuna
Estrutura de Governança

Comitê de direção	Orientação estratégica e governança geral do Fundo
Financiador colaborativo	Fornece informações sobre as principais prioridades ao Comitê Diretor e oferece um fórum para uma colaboração mais profunda com os financiadores.
Painéis de consultoria técnica	Escopo dos processos de financiamento e seleção de propostas, fornecimento de informações sobre estratégias e necessidades específicas do domínio
Secretaria	Gerencia subsídios e relatórios, oferece facilitação, comunicações e suporte operacional

Domínios para conjuntos de dados

O Lacuna Fund fornece recursos para a criação, expansão ou manutenção de conjuntos de dados.

Chamadas atuais para propostas nesses domínios:

- **Processamento de linguagem natural (NLP)**
- **Resistência antimicrobiana (AMR)**

Chamadas anteriores para propostas nesses domínios:

- Agricultura
- Idioma
- Saúde
- Clima

Futuras chamadas para propostas:

- Outras áreas em que o Comitê Diretor identifica uma necessidade

Financiador do NLP 2024

A chamada de propostas do NLP 2024 é possível graças ao generoso apoio do **Google.org**.

Requisitos para a RFP

PNL 2024

NLP 2024 RFP - Objetivo

Objetivo declarado:

O objetivo desta chamada de propostas é apoiar os esforços para desenvolver conjuntos de dados abertos e acessíveis para aplicativos de aprendizado de máquina relacionados ao Processamento de Linguagem Natural (PLN) para **idiomas com poucos recursos na América Latina e na África.**

A capacidade de se comunicar e ser compreendido em seu próprio idioma é fundamental para a inclusão digital e social. As técnicas de processamento de linguagem natural permitiram que os aplicativos de IA facilitassem a inclusão digital e melhorassem a educação, as finanças, a saúde, a agricultura, a comunicação e as respostas a riscos naturais, entre outros. Muitos avanços na PNL fundamental e aplicada resultaram de conjuntos de dados licenciados abertamente e disponíveis ao público.

No entanto, esses conjuntos de dados são **escassos ou inexistentes para muitos idiomas africanos e latino-americanos,** excluindo essas populações dos benefícios da PNL. Muitos modelos atuais de aprendizado de máquina (ML) são informados por conjuntos de dados anglo-cêntricos e/ou traduzidos, sem nuances culturalmente relevantes e criando modelos inutilizáveis para comunidades da América Latina e da África. **Quando existem conjuntos de dados relevantes, eles geralmente se baseiam em textos religiosos ou judiciários do passado, o que leva a uma linguagem desatualizada e tendenciosa. Há necessidade de conjuntos de dados de acesso aberto para facilitar as tecnologias de PNL em idiomas africanos e latino-americanos com poucos recursos** e apoiar o desenvolvimento de conjuntos de dados linguísticos robustos e abrangentes que atendam às necessidades específicas de comunidades sub-representadas.

NLP RFP - Necessidade

O Fundo Lacuna busca propostas de equipes multidisciplinares qualificadas para desenvolver conjuntos de dados de treinamento e avaliação abertos e acessíveis para aplicativos de aprendizado de máquina para PNL em idiomas com poucos recursos e culturas pouco representadas na África e na América Latina. Os conjuntos de dados podem incluir, mas não estão limitados ao seguinte:

- Conjuntos de dados rotulados e não rotulados para tarefas de NLP com poucos recursos
- Corpora de fala
- Conjuntos de dados de tarefas de geração de texto
- Conjuntos de dados multimodais e outros conjuntos de dados inovadores
- - Conjuntos de dados que suportam tarefas com uso intensivo de
- Conjuntos de dados relacionados a corpora de variação dialetal e texto e fala com alternância de código
- Criação ou ampliação de conjuntos de dados de texto e fala específicos do domínio
- Conjuntos de dados que suportam o aprendizado de máquina para linguística
- Em todos os conjuntos de dados: capacidade de resposta ao gênero e inclusão dos principais grupos vulneráveis.

O que estamos procurando

- Coleta e/ou anotação de novos dados;
- Anotação ou liberação de dados existentes;
- Aumentar os conjuntos de dados existentes de diversas fontes para preencher lacunas nos dados locais de verdade, diminuir a parcialidade (como parcialidade geográfica, lacunas de gênero ou outros tipos de parcialidade ou discriminação) ou aumentar a usabilidade dos dados e da tecnologia relacionados à PNL em contextos de baixa e média renda;
- Vincular e harmonizar conjuntos de dados existentes (como entre regiões, tempo, variedades linguísticas, bem como conjuntos de dados específicos do domínio, como dados históricos, de saúde e educação).

Informações sobre a proposta e requisitos de elegibilidade

Informações sobre a proposta

- Informações do candidato
- Narrativa da proposta
- Cronograma e resultados do projeto
- Orçamento
- OBSERVAÇÃO: Oportunidade de mentoria e de formação de parcerias

Requisitos de elegibilidade

Para se qualificar para o financiamento, as organizações devem:

- Ser uma entidade sem fins lucrativos, uma instituição de pesquisa, uma empresa social com fins lucrativos ou uma equipe de tais organizações. As pessoas físicas devem se inscrever por meio de um patrocinador institucional. As parcerias são fortemente incentivadas como forma de fortalecer a colaboração e maximizar os benefícios derivados do uso dos conjuntos de dados, mas somente o candidato principal receberá fundos.
- Ter uma missão de apoio ao bem da sociedade, definida de forma ampla.
- Estar sediado no país ou na região onde os dados serão coletados. **O foco geográfico desta chamada é a África e a América Latina.** Instituições sediadas em outros países ou regiões podem se candidatar como parceiras da instituição líder. Conforme mencionado acima, somente o candidato principal receberá os fundos.

Requisitos de elegibilidade

Para se qualificar para o financiamento, as organizações devem:

- Ter todas as aprovações nacionais ou outras aprovações necessárias para conduzir a pesquisa proposta. O processo de aprovação pode ser conduzido paralelamente à solicitação de subsídio, se necessário. Os custos de aprovação, se houver, são de responsabilidade do candidato.

- Ter a capacidade técnica - ou a capacidade de desenvolver essa capacidade por meio de uma parceria descrita na proposta - para conduzir a rotulagem, a criação, a agregação, a expansão e/ou a manutenção de conjuntos de dados, incluindo a capacidade de aplicar as práticas recomendadas e os padrões estabelecidos no domínio específico (por exemplo, processamento de linguagem natural) para permitir que análises de IA/ML de alta qualidade sejam realizadas por várias entidades.

Principais datas

27 de junho de 2024: Solicitação de propostas aberta

9 de julho de 2024: Webinar para candidatos

12 de julho de 2024: Prazo para perguntas
Envie suas perguntas para secretariat@lacunafund.org

15 de julho de 2024: Prazo final para mentoria

29 de julho de 2024: Respostas publicadas

23 de agosto de 2024: Prazo para apresentação de propostas completas

Primavera de 2025: Subsídios concedidos

Observação: os projetos propostos devem ser concluídos, os conjuntos de dados publicados e os relatórios finais enviados até **outubro de 2026**. Para fins de planejamento, você pode esperar que os acordos sejam concluídos e o trabalho possa começar até **abril de 2025**.

Elementos de propostas sólidas

Elementos de propostas sólidas

- **Equipe multidisciplinar** - experiência em ciência de dados, processamento de linguagem natural (NLP)
- **Casos de uso e envolvimento da comunidade**
- **Declaração** clara **do problema** e como o conjunto de dados ou a agregação proposta ajudará a resolver o problema
- Especificidade sobre o tamanho do conjunto de dados - deve ser de **tamanho e qualidade suficientes** para ser útil em aplicativos futuros
- **As parcerias** são fortemente incentivadas como uma forma de fortalecer a colaboração e maximizar os benefícios derivados do uso dos conjuntos de dados, mas somente o candidato principal receberá os fundos.

Elementos de propostas sólidas

- Considerações **de equidade** (gênero, status socioeconômico, etnia, etc.)
- Consideração e plano para possíveis problemas **de privacidade e ética**
- Planeje o **gerenciamento e o licenciamento de dados**
- Trabalhar em várias áreas sempre que possível e relevante (ou seja, regiões, tempo, variedades linguísticas)
- **O orçamento** é apropriado para o tamanho do conjunto de dados produzido

Envolvimento da comunidade

- Descreva a consulta anterior e/ou a colaboração proposta com os beneficiários pretendidos.
  - Quando e onde você se reuniu/se reunirá com os parceiros
  - Como os parceiros se envolverão:
    - Identificação das necessidades de dados
    - Coleta de dados, rotulagem
    - Governança de dados
    - Uso do conjunto de dados
  - Como os parceiros se beneficiarão do conjunto de dados novo/ampliado

Privacidade e ética

- Explique como sua equipe abordará o assunto:

- a) preocupações com a privacidade,
- b) potencial de uso indevido no downstream,
- c) possíveis vetores de discriminação (por exemplo, gênero) e
- d) condições de trabalho justas e equitativas, se houver rotuladores pagos envolvidos no projeto.

- Descreva o processo que você usará para examinar possíveis problemas éticos (por exemplo, um conselho de revisão institucional, etc.).

Plano de sustentabilidade e comunicações

- Descreva como o conjunto de dados será mantido e/ou expandido além do financiamento inicial (por exemplo, por meio de um modelo de linha de base, aplicativos de ML resultantes, por uma comunidade dedicada ou um grupo de partes interessadas com um modelo de governança robusto para o conjunto de dados aberto) e como um caso de uso em potencial poderia sustentar o projeto.

- Descreva as atividades de comunicação para divulgar o(s) conjunto(s) de dados. Essas atividades podem incluir atividades de rede com possíveis usuários de dados; apresentação do(s) conjunto(s) de dados em uma conferência; organização de um workshop sobre sustentabilidade do conjunto de dados com as partes interessadas; ou estabelecimento de um comitê de sustentabilidade.

Padrões e compartilhamento de dados

- Localizável - fácil de encontrar em uma plataforma pública e amplamente utilizada
- Acessível - formato aberto ((CCBY 4.0 or CCBY SA 4.0)
- Interoperável - formato de dados
- Reutilizável - metadados

- Sustentável - plano de manutenção
- Compartilhado - comunidade de usuários engajada e plano para compartilhar o conjunto de dados depois de concluído

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

Lacuna Fund Intellectual Property Policy: IP-Policy_LacunaFund.pdf

Orçamento e custos permitidos

- Forneça um orçamento para a conclusão do conjunto de dados proposto enviado por meio do portal SurveyMonkey Apply. Esse orçamento deve ser formatado no modelo de orçamento do Lacuna Fund, disponível no portal do candidato.

O total disponível é de aproximadamente US$ 1 milhão. Gostaríamos de financiar projetos em cada uma das regiões-alvo (África, América Latina) e prevemos apoiar de 6 a 8 projetos menores com orçamentos de até US$ 100 mil e de 2 a 3 projetos maiores e mais complexos com orçamentos que variam de US$ 100 a US$ 250 mil. O Painel Técnico Consultivo avaliará a viabilidade e a adequação do orçamento, bem como a ligação entre o orçamento e a narrativa do subsídio como parte dos critérios de seleção.

Orçamento e custos permitidos

Os orçamentos podem incluir, mas não estão limitados a, custos para:

- Capacitação relacionada à coleta de dados e garantia de qualidade/controle de qualidade;
- Coleta de dados (incluindo compensação justa pelo fornecimento de dados);
- Rotulagem de dados (incluindo compensação justa pela rotulagem de dados);
- QA/QC ou verificação;
- Pós-processamento de dados;
- Publicação de dados;
- Modelo de linha de base
- Licenciamento.
- Publicação dos resultados em acesso aberto.
- É hora de preparar uma declaração de dados para o conjunto de dados.
- Esforços de crowdsourcing, como os "label-a-thons".
- Armazenamento de dados.
- Poder de computação.
- Oficina
- Atividades de comunicação, incluindo a participação em conferências para até dois eventos de apresentação dos conjuntos de dados

Considerações sobre a qualidade dos dados

Hospedagem e acessibilidade - Princípios de Lacuna

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

Hospedagem - Diretrizes da Lacuna

- Atribui um identificador de objeto digital (DOI) para conjuntos de dados ou permite que um seja anexado como parte dos metadados.
- É indexado pelos principais mecanismos de pesquisa (por exemplo, Google Dataset Search ou ferramentas semelhantes).
- É confiável e persistente

É ótimo ter:

- Quantifica o número de visualizações e downloads da página de destino para o conjunto de dados.
- Coleta informações de contato para downloads de conjuntos de dados de uma forma que maximiza a conversão.

Hospedagem de conjuntos de dados

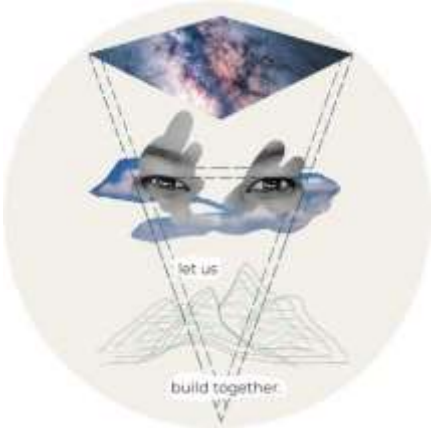
A documentação e a hospedagem propostas estão alinhadas com o Guia de documentação e hospedagem de conjuntos de dados do Lacuna Fund.

O Lacuna Fund solicita aos beneficiários que incluam a seguinte documentação quando os conjuntos de dados forem enviados:

- Arquivo de metadados
- Folha de dados
- Identificador de objeto digital (DOI)

Oportunidade de mentoria: O Programa de Mentoria Masakhane Oportunidade:

Fundação de Pesquisa Masakhane



O Programa de Mentoria Masakhane

- **Missão:** A MRF é uma organização de base cuja missão é fortalecer e estimular a pesquisa de PNL em idiomas africanos, para africanos e por africanos. O objetivo do MRF é que os africanos moldem e se apropriem desses avanços tecnológicos rumo à dignidade humana, ao bem-estar e à equidade, por meio da construção de comunidades inclusivas, da pesquisa participativa aberta e da multidisciplinaridade.

- A MRF oferecerá aos candidatos da África e da América Latina uma oportunidade especial de se juntarem à comunidade MRF e de encontrarem um mentor para aqueles que estiverem interessados, que analisará o esboço da proposta e discutirá as possibilidades de fortalecê-la.

O Programa de Mentoria Masakhane

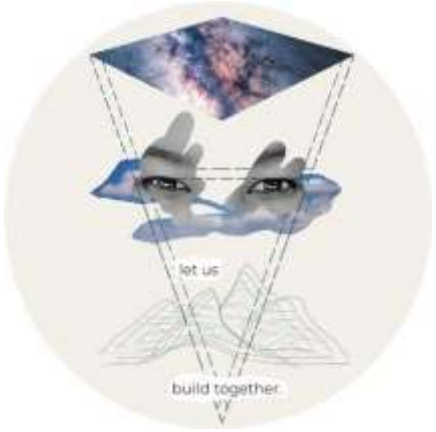
- Inscreva-se no **Formulário do Google** abaixo https://forms.gle/8u5GobjKXYo7gujt5
- Forneça uma breve descrição (resumo de 250 palavras) da proposta do conjunto de dados
- Os mentorandos podem solicitar diferentes formas de assistência, como
    - "Discutindo lacunas na PNL com poucos recursos",
    - "escrever uma proposta de pesquisa", e
    - "preparação de orçamentos".

O Programa de Mentoria Masakhane

- Os candidatos interessados são incentivados a se inscrever para uma sessão de orientação pelo menos seis semanas antes da data de vencimento da proposta, ou seja, até **15 de julho de 2024**

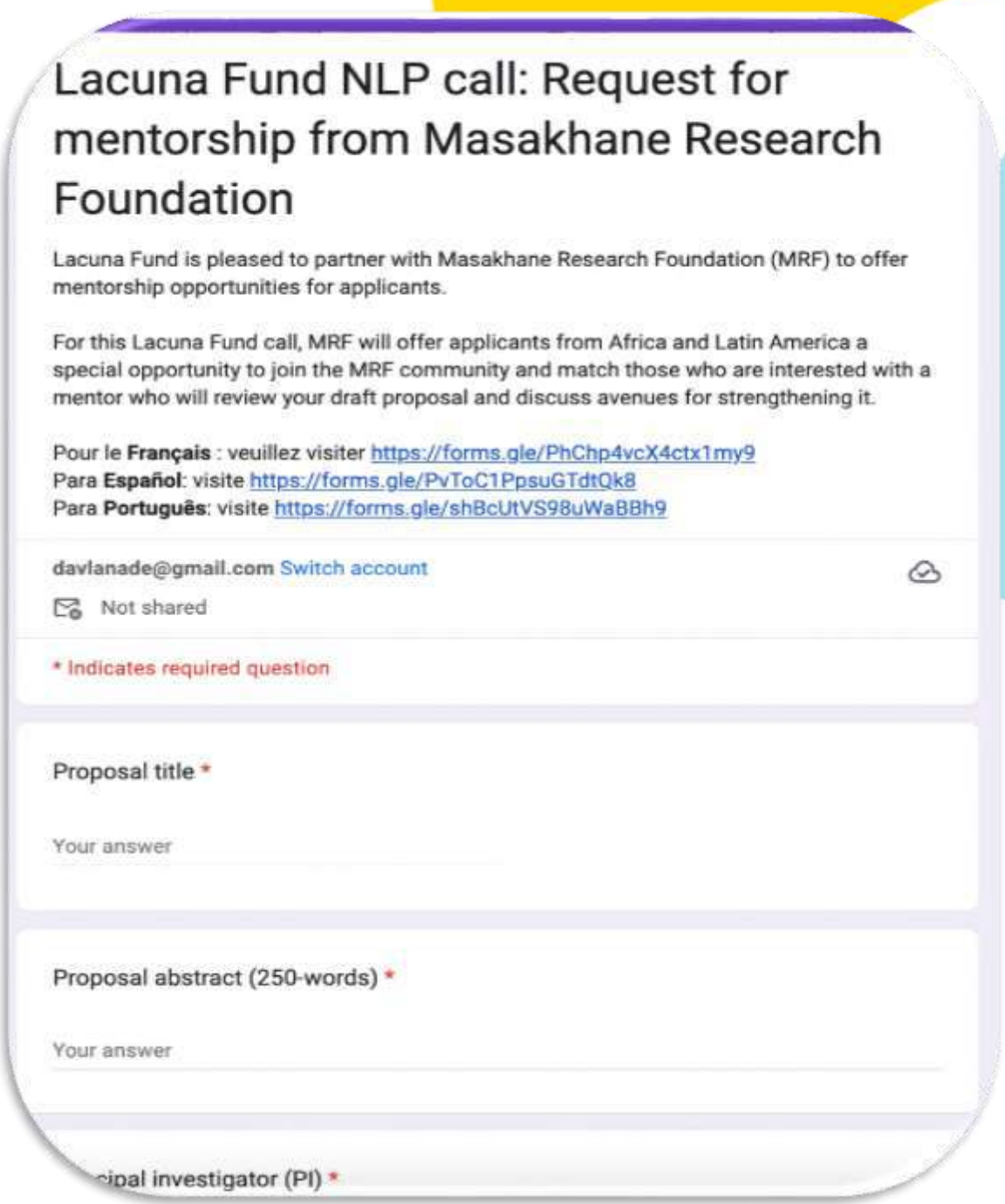
- Os mentores serão designados por ordem de chegada.

- Espera-se que todos os candidatos leiam e cumpram o **código de ética e conduta** do programa de mentoria.



Para obter mais informações, acesse:
https://lacunafund.org/apply/

Dúvidas?
E-mail:
masakhane_leadership@googlegroups.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

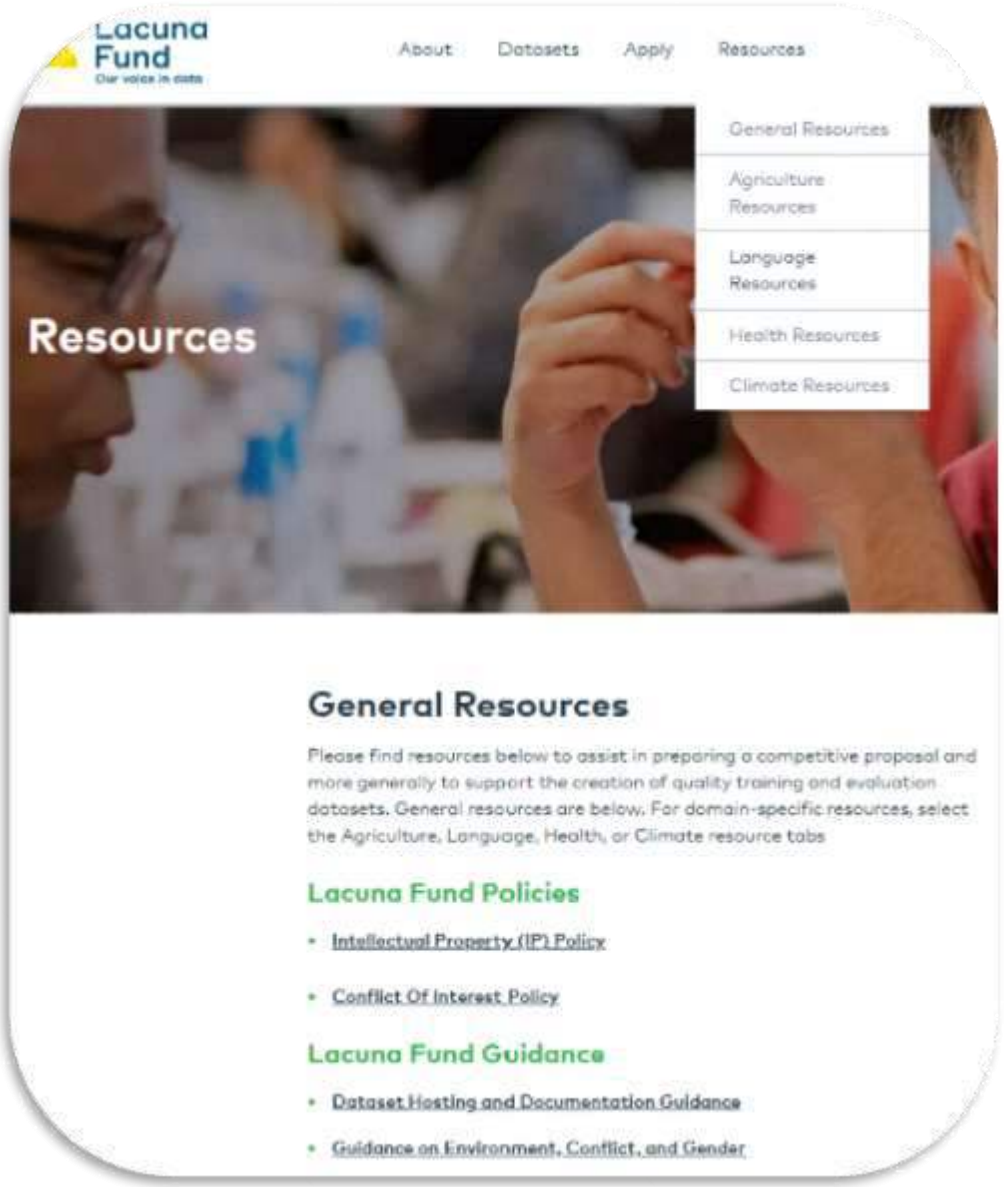
Proposal abstract (250-words) *

Your answer

cipal investigator (PI) *

Recursos

Recursos



O site do Lacuna Fund inclui vários recursos, como referências relevantes sobre **qualidade de dados e documentação** para ajudar os candidatos a preparar uma candidatura competitiva.

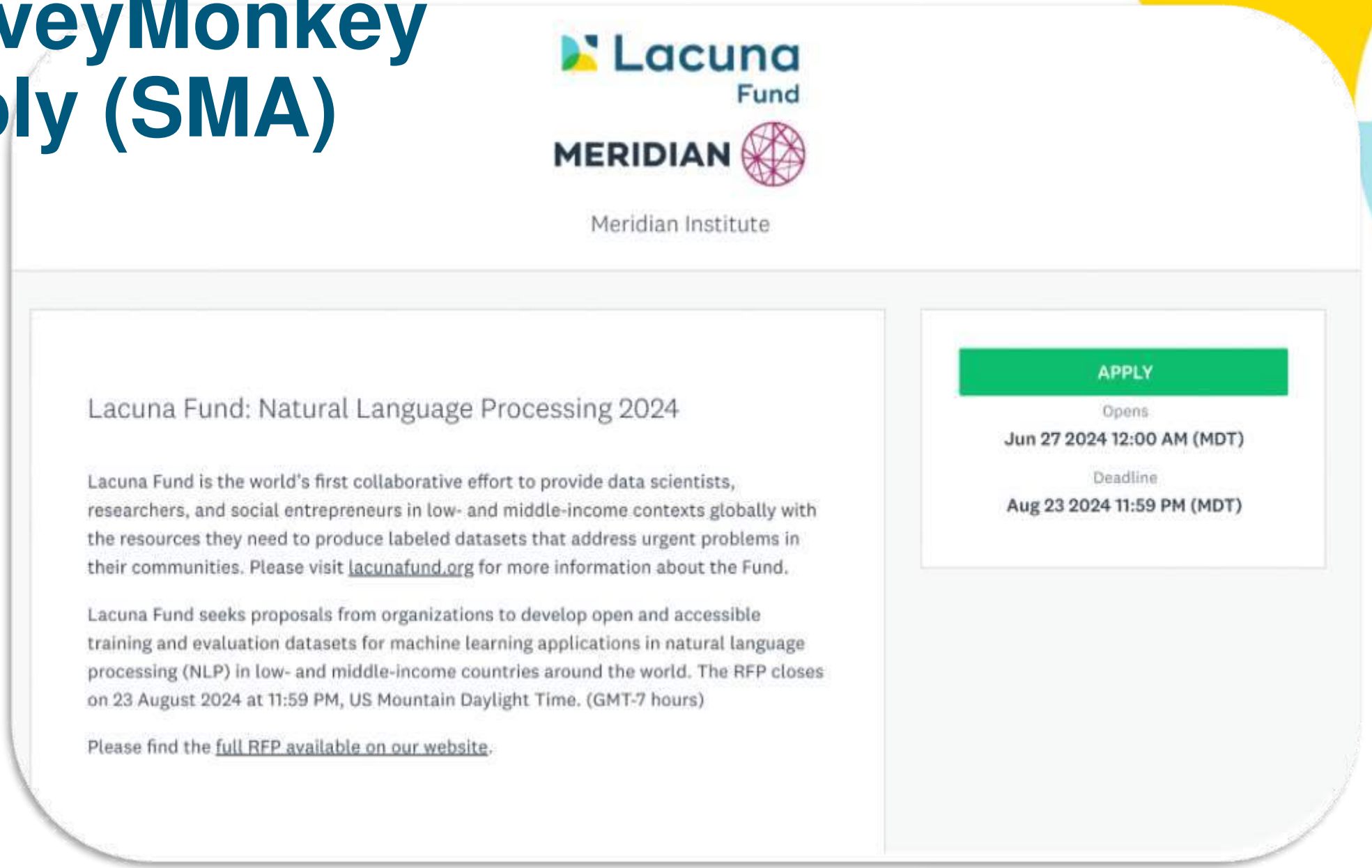
Envio de proposta

Os envios de propostas só serão aceitos por meio do portal de inscrição SurveyMonkey Apply, disponível em www.lacunafund.org/apply.

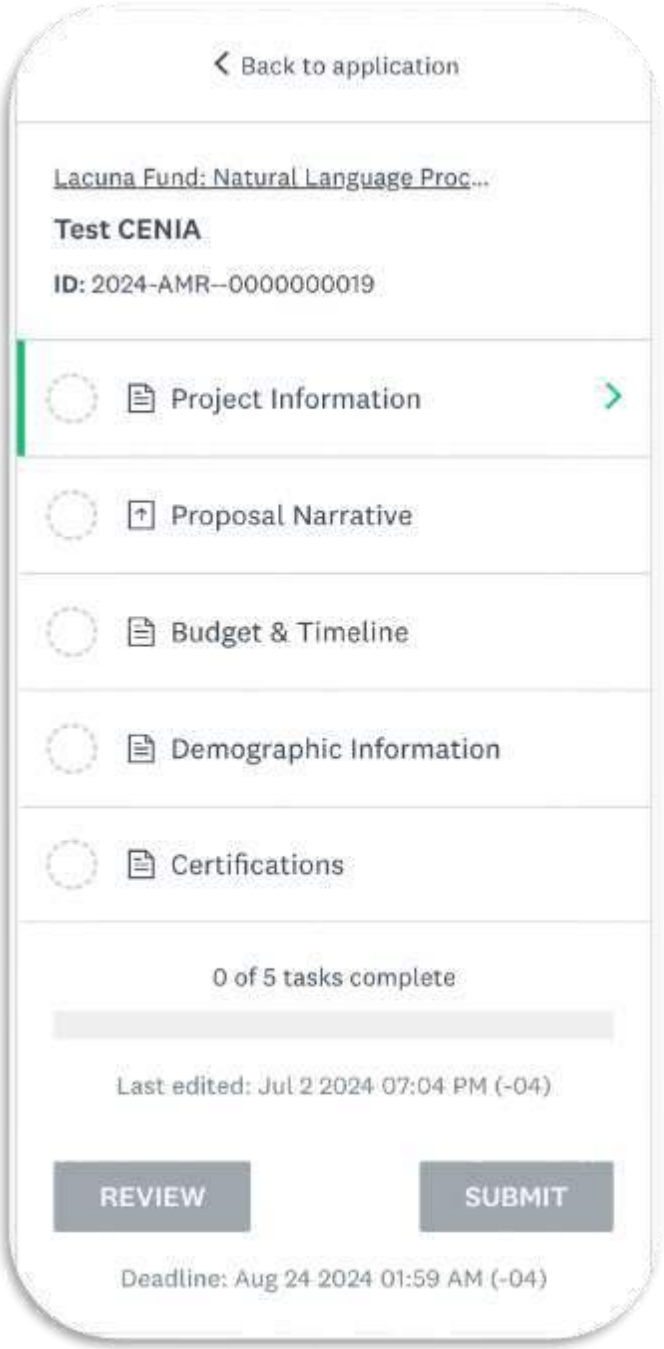
As inscrições podem ser enviadas em **inglês**, **francês**, **português** e **espanhol**.

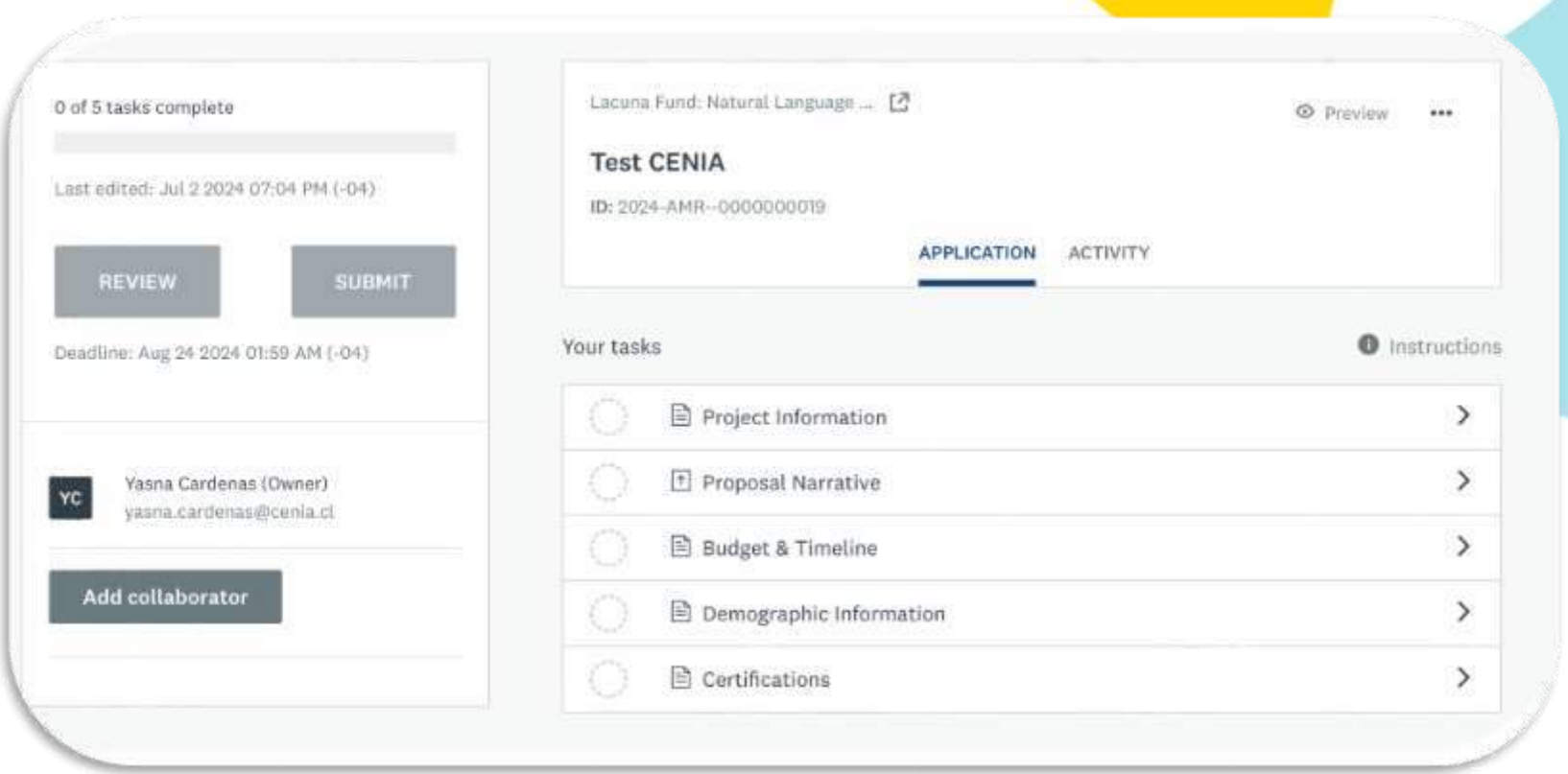
Observação: Selecione o idioma desejado usando a guia suspensa no portal de inscrição. ***Para aqueles que enviarem uma inscrição em português, você pode se inscrever usando os portais em inglês, espanhol ou francês. No momento, não temos uma opção de portal em português disponível.*** *No entanto, as propostas enviadas em português em qualquer portal são aceitas e serão analisadas*

Portal SurveyMonkey Apply (SMA)



Aplicativo


Back to application
Lacuna Fund: Natural Language Proc...
Test CENIA
ID: 2024-AMR--0000000019
Project Information
Proposal Narrative
Budget & Timeline
Demographic Information
Certifications
0 of 5 tasks complete
Last edited: Jul 2 2024 07:04 PM (-04)
REVIEW
SUBMIT
Deadline: Aug 24 2024 01:59 AM (-04)


0 of 5 tasks complete
Last edited: Jul 2 2024 07:04 PM (-04)
REVIEW
SUBMIT
Deadline: Aug 24 2024 01:59 AM (-04)
YC
Yasna Cardenas (Owner)
yasna.cardenas@cenia.cl
Add collaborator
Lacuna Fund: Natural Language ...
Preview
Test CENIA
ID: 2024-AMR--0000000019
APPLICATION
ACTIVITY
Your tasks
Instructions
Project Information
Proposal Narrative
Budget & Timeline
Demographic Information
Certifications

Dúvidas?

Próximas etapas

- **Envie perguntas adicionais para secretariat@lacunafund.org até 12 de julho de 2024.**
- **Respostas postadas** publicamente na página Lacuna Fund Apply em **29 de julho de 2024**
- **As propostas devem ser entregues em 23 de agosto de 2024.**