Conjuntos de datos inclusivos para el aprendizaje automátizado

Procesamiento del Lenguaje Natural 2024
Seminario Web para postulantes

9 de julio

Lacuna Fund

Fondo Lacuna Hub

CENIA

CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL

Gestión de la convocatoria de propuestas del PNL 2024

Con sede en Chile

ALVARO SOTO
CENIA

CRISTINA FLORES
CENIA

Oradores

KATRINA GEHMAN

INSTITUTO MERIDIAN, SECRETARÍA DEL FONDO LACUNA

DR. ALBERT KAHIRA, ASESOR EN CALIDAD DE DATOS, DATAWISE ÁFRICA

TUTORÍA Y MATCHMAKING, MASAKHANE

Miembros de la Comisión Asesora Técnica (TAP)

AUDREY JULIA WALEGHWA MBOGHO

USIU-AFRICA

FELIPE DANIEL HASLER SANDOVAL, PHD

UNIVERSIDAD DE CHILE

Objetivos de la reunión

NLP 2024

- Entregar a los posibles postulantes información sobre el Fondo Lacuna y los requisitos de las propuestas.

- Dejar tiempo para las preguntas de los postulantes sobre la solicitud de propuestas (RFP).

Agenda

00:00	Bienvenida, presentación, examen del orden del día
00:10	Presentación: Presentación del Fondo Lacuna y requisitos para las solicitudes de propuestas
00:40	Presentación: Consideraciones sobre la calidad de datos y Hosting
00:50	Presentación: Oportunidad de tutoría y establecimiento de contactos (*matchmaking)*
01:00	Preguntas y respuestas
01:25	Próximos pasos
01:30	Se levanta la sesión

Una laguna es un espacio o una parte faltante de manuscrito.

El Fondo Lacuna apoya la creación, ampliación y mantenimiento de conjuntos de datos de formación y evaluación equitativa que permitan a las herramientas de aprendizaje automático abordar mejor los problemas urgentes en contextos de ingresos bajos y medios en todo el mundo.

Principios que guían el Fondo Lacuna:

- Accesibilidad
- Equidad
- Ética
- Enfoque participativo
- Calidad
- Impacto transformador

Patrocinadores

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

Fondo Lacuna
Estructura Administrativa

Comité Directivo	Orientación estratégica y gobernanza general del Fondo
Patrocinador Colaborador	Proporciona información sobre las prioridades clave al Comité Directivo y constituye un foro para una mayor colaboración de los patrocinadores-
Paneles de asesoramiento técnico	Analiza los procesos de financiamiento y selecciona las propuestas, aporta información sobre las estrategias y necesidades específicas de cada ámbito.
Secretaría	Gestiona las subvenciones y los informes, proporciona facilitación, comunicaciones y apoyo operativo.

Dominios para conjuntos de datos

El Fondo Lacuna proporciona recursos para la creación, ampliación o mantenimiento de conjuntos de datos.

Convocatorias de propuestas actuales en estos ámbitos:

- **Procesamiento del lenguaje natural (PLN)**
- **Resistencia antimicrobiana (RAM)**

Convocatorias de propuestas anteriores en estos ámbitos:

- Agricultura
- Idioma
- Salud
- Clima

Futuras convocatorias de propuestas:

- Otros ámbitos en los que el Comité Directivo detecta una necesidad

Patrocinador del PNL 2024

La convocatoria de propuestas de PNL 2024 es posible gracias al generoso apoyo de **Google.org**.

Requisitos para la RFP

PNL 2024

PNL 2024 RFP - Objetivo

Objetivo declarado:

El objetivo de esta convocatoria de propuestas es apoyar los esfuerzos para desarrollar conjuntos de datos abiertos y accesibles para aplicaciones de aprendizaje automático relacionadas con el Procesamiento del Lenguaje Natural (PLN) para **lenguas en regiones de bajos recursos en América Latina y África.**

La capacidad de comunicarse y hacerse entender en la propia lengua es fundamental para la inclusión digital y social. Las técnicas de procesamiento del lenguaje natural han permitido aplicaciones de IA que facilitan la inclusión digital y mejoras en la educación, finanzas, sanidad, en agricultura, comunicación y respuestas a los riesgos naturales, entre otros. Muchos avances en PNL, tanto fundamentales como aplicados, se han derivado de conjuntos de datos con licencia abierta y a disposición del público.

Sin embargo, estos conjuntos de datos son **escasos o inexistentes para muchas lenguas africanas y latinoamericanas, lo que** excluye a estas poblaciones de los beneficios del PLN. Muchos modelos actuales de aprendizaje automático se basan en conjuntos de datos anglo céntricos o traducidos, por lo que carecen de matices culturalmente relevantes y crean modelos inutilizables para las comunidades de América Latina y África. **Cuando existen conjuntos de datos relevantes, a menudo se basan en textos religiosos o judiciales del pasado, lo que da lugar a un lenguaje anticuado y sesgado. Se necesitan conjuntos de datos de libre acceso para facilitar el uso de tecnologías de PNL en lenguas africanas y latinoamericanas en regiones de escasos recursos** y apoyar el desarrollo de conjuntos de datos lingüísticos sólidos y exhaustivos que respondan a las necesidades específicas de las comunidades infrarrepresentadas.

RFP de NLP - Necesidad

El Fondo Lacuna busca propuestas de equipos cualificados y multidisciplinarios para desarrollar conjuntos de datos de capacitación y evaluación abiertos y accesibles para aplicaciones de aprendizaje automático para PNL en lenguas de regiones de bajos recursos y culturas subrepresentadas en África y América Latina. Los conjuntos de datos pueden incluir, entre otros, los siguientes:

- Conjuntos de datos etiquetados y sin etiquetar para tareas de PLN con pocos recursos
- Corpus lingüísticos
- Conjuntos de datos de tareas de generación de texto
- Conjuntos de datos multimodales y otros datos innovadores
- - Conjuntos de datos para tareas intensivas en conocimiento
- Conjuntos de datos relacionados con corpus de variación dialéctica y texto y habla codificados
- Creación o aumento de conjuntos de datos de texto y voz específicos de un dominio
- Conjuntos de datos de aprendizaje automático para lingüística
- En todos los conjuntos de datos: perspectiva de género e inclusión de los principales grupos vulnerables.

Qué buscamos

- Recopilación y/o anotación de nuevos datos;
- Anotar o liberar datos existentes;
- Aumentar los conjuntos de datos existentes procedentes de diversas fuentes para completar lagunas en los datos locales de la verdad sobre el terreno, disminuir los sesgos (como los geográficos, los de género u otros tipos de sesgo o discriminación), o aumentar la usabilidad de los datos y la tecnología relacionados con la PNL en contextos de ingresos bajos y medios;
- Vincular y armonizar los conjuntos de datos existentes (por ejemplo, entre regiones, épocas, variedades lingüísticas, así como conjuntos de datos específicos de un ámbito, como datos históricos, sanitarios y educativos).

Información sobre la propuesta y requisitos de admisibilidad

Información sobre la propuesta

- Información del solicitante
- Narrativa de la propuesta
- Calendario y resultados del proyecto
- Presupuesto
- NOTA: Oportunidad de tutoría y establecimiento de contactos *(matchmaking*)

Requisitos

Para poder optar a la financiación, las organizaciones deben

- Ser una entidad sin fines de lucro, una institución de investigación, una empresa social con fines de lucro o un equipo de estas organizaciones. Las personas físicas deben presentar su solicitud a través de un patrocinador institucional. Se recomienda encarecidamente la asociación para reforzar la colaboración y maximizar los beneficios derivados del uso de los conjuntos de datos, pero sólo el solicitante principal recibirá fondos.
- Tener una misión de apoyo al bien social, definido en sentido amplio.
- Tener su sede en el país o región donde se recopilarán los datos. **Esta convocatoria se centra en África y América Latina.** Las instituciones con sede en otros países o regiones pueden presentar su candidatura como socios de la institución principal. Como ya se ha indicado, sólo el postulante principal recibirá fondos.

Requisitos

Para poder optar a financiamiento, las organizaciones deben

- Contar con todas las autorizaciones nacionales o de otro tipo necesarias para llevar a cabo la investigación propuesta. En caso necesario, el proceso de aprobación podrá realizarse paralelamente a la solicitud de subvención. Los costes de aprobación, en su caso, correrán a cargo del postulqnte.

- Tener la capacidad técnica -o la capacidad de crear esta capacidad a través de una asociación descrita en la propuesta- para llevar a cabo el etiquetado, la creación, la agregación, la expansión y/o el mantenimiento del conjunto de datos, incluida la capacidad de aplicar las mejores prácticas y las normas establecidas en el ámbito específico (por ejemplo, el procesamiento del lenguaje natural) para permitir que múltiples entidades realicen análisis de IA/ML de alta calidad.

Fechas clave

27 de junio de 2024: Solicitud de propuestas abierta

9 de julio de 2024: Seminario Webinar para postulantes

12 de julio de 2024: Plazo de preguntas
Envíe sus preguntas a secretariat@lacunafund.org

15 de julio de 2024: Fecha límite de tutoría

29 de julio de 2024: Respuestas publicadas

23 de agosto de 2024: Presentación de propuestas completas

Primavera de 2025: Subvenciones concedidas

Primavera 2025 Nota: Adjudicación de subvenciones: Los proyectos propuestos deben estar terminados, los conjuntos de datos publicados y los informes finales presentados en **octubre de 2026**. A efectos de planificación, cabe esperar que los acuerdos estén concluidos y los trabajos puedan comenzar en **abril de 2025**.

Elementos parte de una sólida propuesta

Elementos de una propuesta sólida

- **Equipo multidisciplinario** - experiencia en ciencia de datos, Procesamiento del Lenguaje Natural (PLN)
- **Casos prácticos y participación ciudadana**
- **Planteamiento** claro **del** problema y cómo el conjunto de datos o la agregación propuestos ayudarán a resolverlo.
- Especificidad sobre el tamaño del conjunto de datos: debe tener un **tamaño y calidad suficientes** para ser útil en futuras aplicaciones.
- Se recomienda encarecidamente la **creación de asociaciones para** reforzar la colaboración y maximizar los beneficios derivados del uso de los conjuntos de datos, pero sólo el postulante principal recibirá fondos.

Elementos de una propuesta sólida

- Consideraciones de **equidad** (género, estatus socioeconómico, etnia, etc.)
- Consideración y plan de posibles problemas en cuanto a **privacidad y ética**
- Plan de **gestión de datos y licencias**
- Trabajar en distintos ámbitos siempre que sea posible y pertinente (por ejemplo, regiones, tiempo, variedades lingüísticas).
- **El presupuesto** es adecuado para el tamaño del conjunto de datos producido

Participación comunitaria

- Describa las consultas previas y/o la colaboración propuesta con los beneficiarios previstos.
  - Cuándo y dónde se ha reunido/se reunirá con los socios
  - Cómo participarán los socios en:
    - Determinandode las necesidades de datos
    - Recogida de datos, etiquetado
    - Gobernanza de datos
    - Uso del conjunto de datos
  - Cómo se beneficiarán los socios del conjunto de datos nuevo/ampliado

Privacidad y ética

- Explique cómo abordará su equipo:

- a) cuestiones de privacidad,
- b) potencial de uso indebido posterior,
- c) posibles vectores de discriminación (por ejemplo, el sexo), y
- d) condiciones de trabajo justas y equitativas, si en el proyecto participan etiquetadores remunerados.

- Describa el proceso que utilizará para detectar posibles problemas éticos (por ejemplo, una junta de revisión institucional, etc.).

Plan de Sostenibilidad y Comunicación

- Describa cómo se mantendrá y/o ampliará el conjunto de datos más allá del financiamiento inicial (por ejemplo, a través de un modelo de referencia, aplicaciones de ML resultantes, por una comunidad dedicada o un grupo de partes interesadas con un modelo de gobernanza sólido para el conjunto de datos abierto) y cómo un caso de uso potencial podría sostener el proyecto.

- Esbozar actividades de comunicación para dar a conocer el conjunto o conjuntos de datos. Podrían incluir actividades de creación de redes con posibles usuarios de los datos, la presentación de los conjuntos de datos en conferencias, la organización de un taller sobre sostenibilidad de los conjuntos de datos con las partes interesadas o la creación de un comité de sostenibilidad.

Normas e intercambio de datos

- Localizable: fácil de encontrar en una plataforma pública y ampliamente utilizada.
- Accesible - formato abierto (CCBY 4.0 o CCBY SA 4.0)
- Interoperabilidad - formato de los datos
- Reutilizables - metadatos

- Sostenible: plan de mantenimiento
- Compartidos: comunidad de usuarios comprometida y plan para compartir el conjunto de datos una vez completado.

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

Lacuna Fund Intellectual Property Policy: IP-Policy_LacunaFund.pdf

Presupuesto y costos autorizados

- Proporcione un presupuesto para la realización del conjunto de datos propuesto presentado a través del portal SurveyMonkey Apply. El formato debe ser el de la plantilla de presupuesto del Fondo Lacuna, disponible en el portal del postulante.

La dotación total disponible es de aproximadamente un millón de dólares estadounidenses. Nos gustaría financiar proyectos en cada una de las regiones objetivo (África, América Latina) y prevemos apoyar entre 6 y 8 proyectos más pequeños con presupuestos de hasta 100.000 USD y entre 2 y 3 proyectos más grandes y complejos con presupuestos de entre 100.000 y 250.000 USD. El Panel de Asesoramiento Técnico evaluará la viabilidad y adecuación del presupuesto, así como la relación entre el presupuesto y la descripción de la subvención como parte de los criterios de selección.

Presupuesto y costes autorizados

Los presupuestos pueden incluir, entre otros, los costes de:

- Desarrollo de capacidades relacionadas con la recopilación de datos y la garantía y el control de calidad;
- Recogida de datos (incluida una compensación justa por el suministro de datos);
- Etiquetado de datos (incluida una compensación justa por el etiquetado de datos);
- Control de calidad o verificación;
- Tratamiento posterior de los datos;
- Publicación de datos;
- Modelo de referencia
- Concesión de licencias.
- Publicación de los resultados en acceso abierto.
- Es hora de preparar una declaración de datos para el conjunto de datos.
- Esfuerzos de *crowdsourcing*, como las maratones de etiquetado.
- Almacenamiento de datos.
- Ponderación de cálculo.
- Taller
- Actividades de comunicación, incluida la asistencia a un máximo de dos conferencias para presentar los conjuntos de datos.

Consideraciones sobre calidad de los datos

Hosting y accesibilidad

Principios de Lacuna

Hosting y accesibilidad -Principios de Lacuna

Accesibilidad, Equidad, Ética, Enfoque Participativo, Calidad, Impacto Transformacional

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

Hosting - Directrices Lacuna

- Asigna un identificador de objeto digital (DOI) a los conjuntos de datos o permite adjuntar uno como parte de los metadatos.

- Está indexado por los principales motores de búsqueda (por ejemplo, Google Dataset Search o herramientas similares).

- Es fiable y robusto

Es genial tenerlo:

- Cuantifica el número de visitas y descargas de la página de destino para el conjunto de datos.

- Recopila información de contacto para descargas de conjuntos de datos de forma que se maximice la conversión.

Hosting de conjuntos de datos

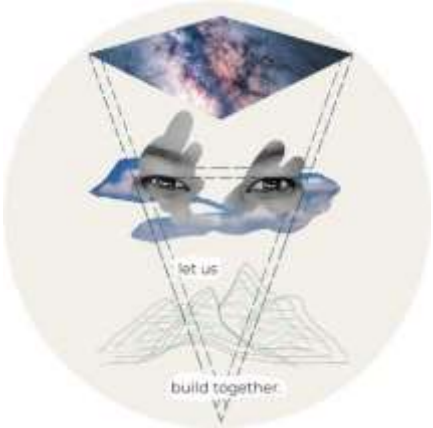
La documentación y el Hosting propuestos se ajustan a la Guía de documentación y Hosting de conjuntos de datos del Fondo Lacuna.

El Fondo Lacuna pide a los beneficiarios que incluyan la siguiente documentación cuando presenten los conjuntos de datos:

- Archivo de metadatos
- Ficha de datos
- Identificador digital de objetos (DOI)

Oportunidad de tutoría: El Programa de Mentores “Masakhane Oportunidad”:

Fundación de Investigación Masakhane



Programa de tutoría Masakhane

- **Misión:** MRF es una organización de base cuya misión es fortalecer y estimular la investigación en PNL en lenguas africanas, para africanos y por africanos. El objetivo de MRF es que los africanos den forma y se apropien de estos avances tecnológicos hacia la dignidad humana, el bienestar y la equidad, a través de la construcción de comunidades inclusivas, la investigación participativa abierta y la multidisciplinariedad.

- La MRF ofrecerá a los postulantes de África y América Latina una oportunidad especial de unirse a la comunidad MRF y pondrá en contacto a los interesados con un mentor que revisará el borrador de su propuesta y discutirá las vías para reforzarla.

Programa de tutoría Masakhane

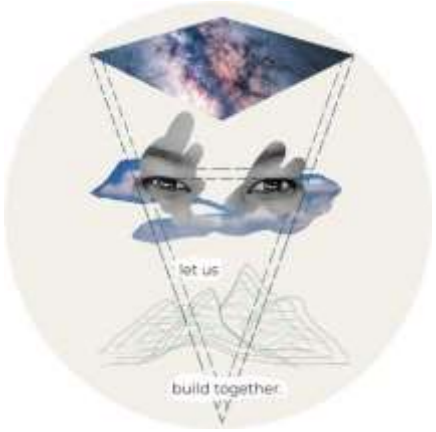
- Envíe su solicitud a través del **formulario de Google** https://forms.gle/8u5GobjKXYo7gujt5
- Proporcione una breve descripción (resumen de 250 palabras) de la propuesta de conjunto de datos
- Los alumnos pueden solicitar distintas formas de ayuda, como por ejemplo
    - "Discutir las lagunas de la PNL con pocos recursos",
    - "redactar una propuesta de investigación", y
    - "preparar los presupuestos".

Programa de tutoría Masakhane

- Se anima a los candidatos interesados en solicitar una sesión de tutoría al menos 6 semanas antes de la fecha límite de presentación de propuestas, el **15 de julio de 2024.**

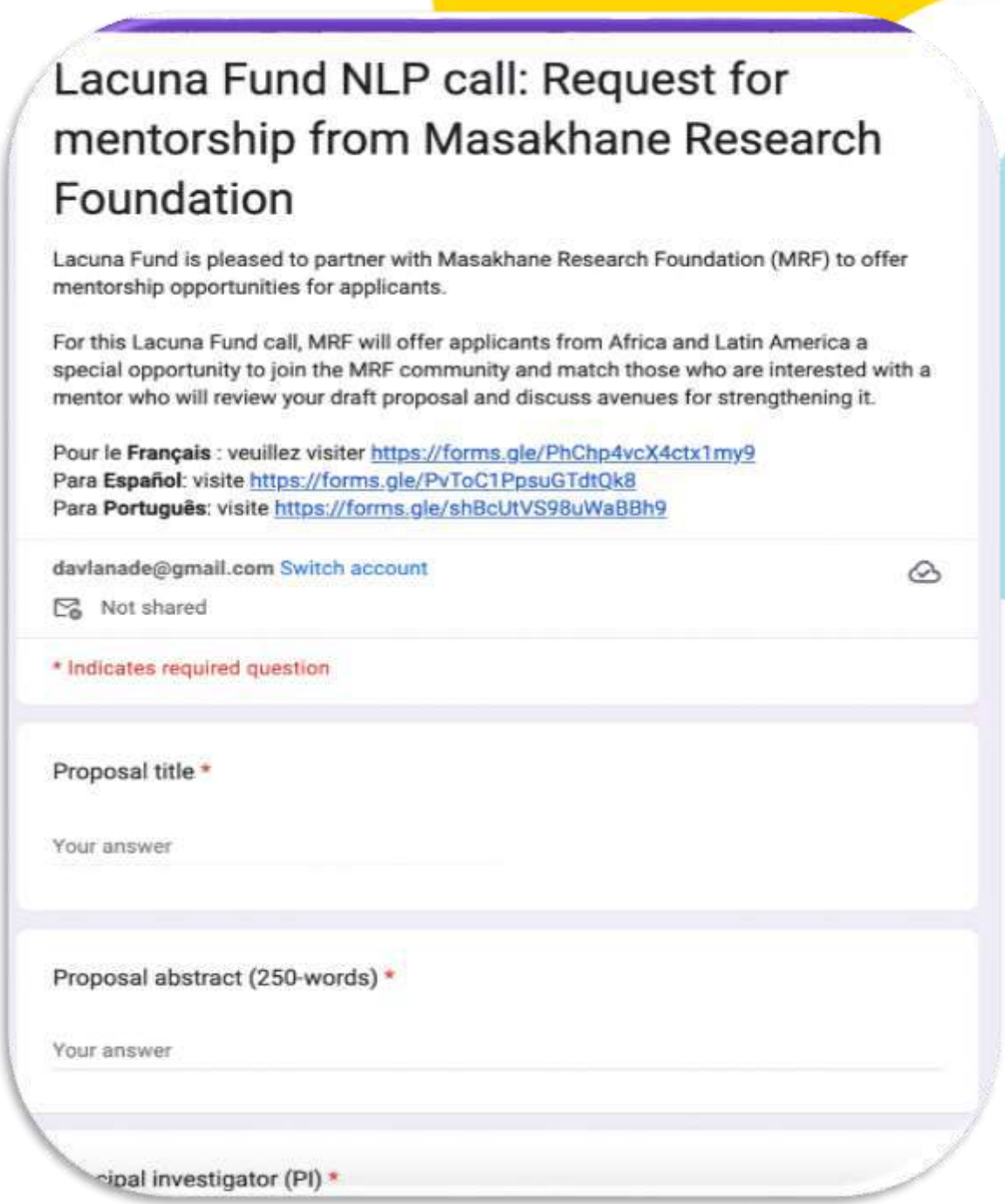
- Los mentores se asignarán por orden de llegada.

- Todos los postulantesdeben leer y respetar el **código ético y de conducta del** programa de tutoría.



Más información en:
https://lacunafund.org/apply/

¿Preguntas?
Correo electrónico:
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

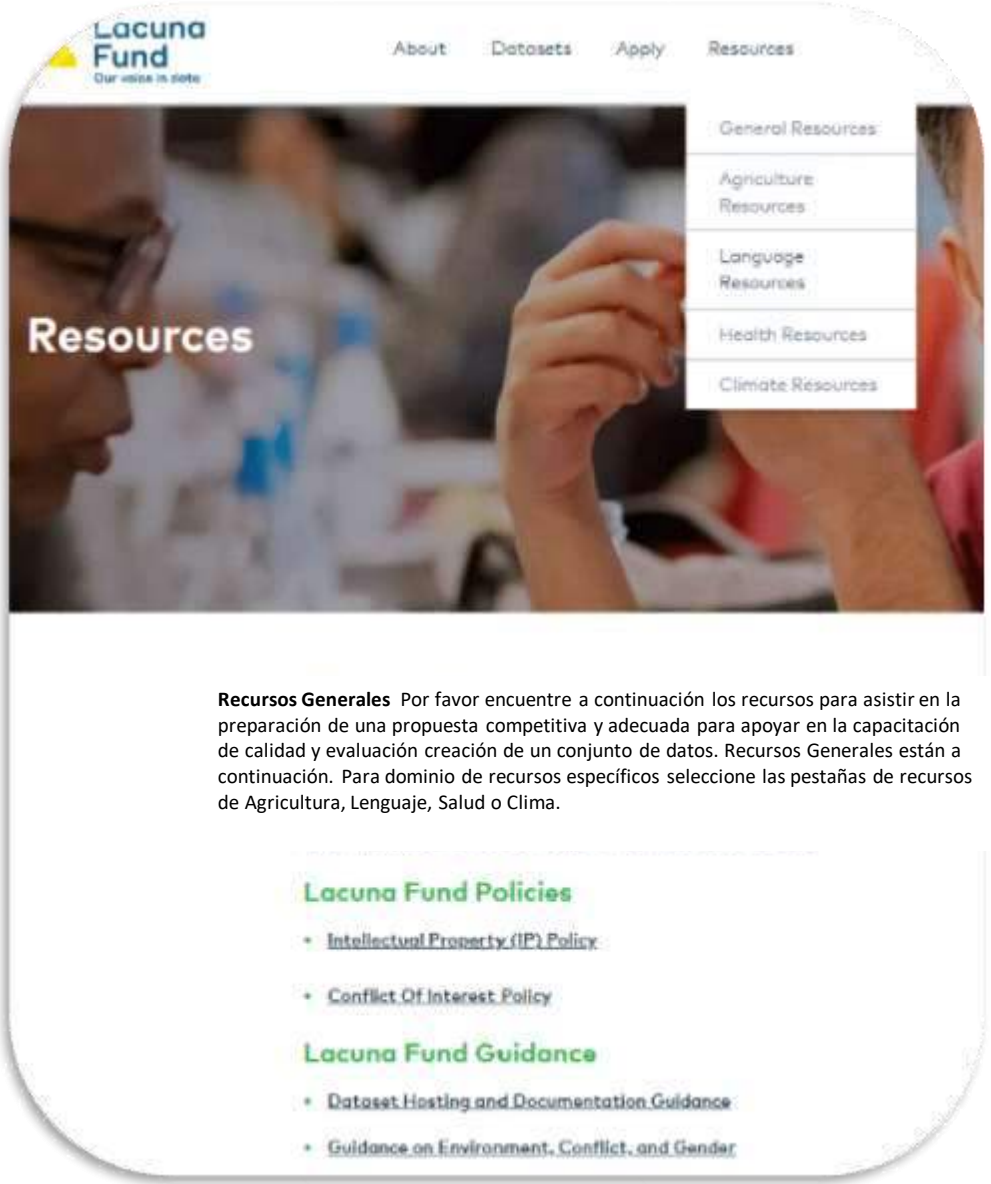
Proposal abstract (250-words) *

Your answer

cipal investigator (PI) *

Recursos

Recursos



El sitio web del Fondo Lacuna incluye diversos recursos, como referencias sobre la **calidad de los datos y la documentación** para ayudar a los postulantesa preparar una solicitud competitiva.

Presentación de propuestas

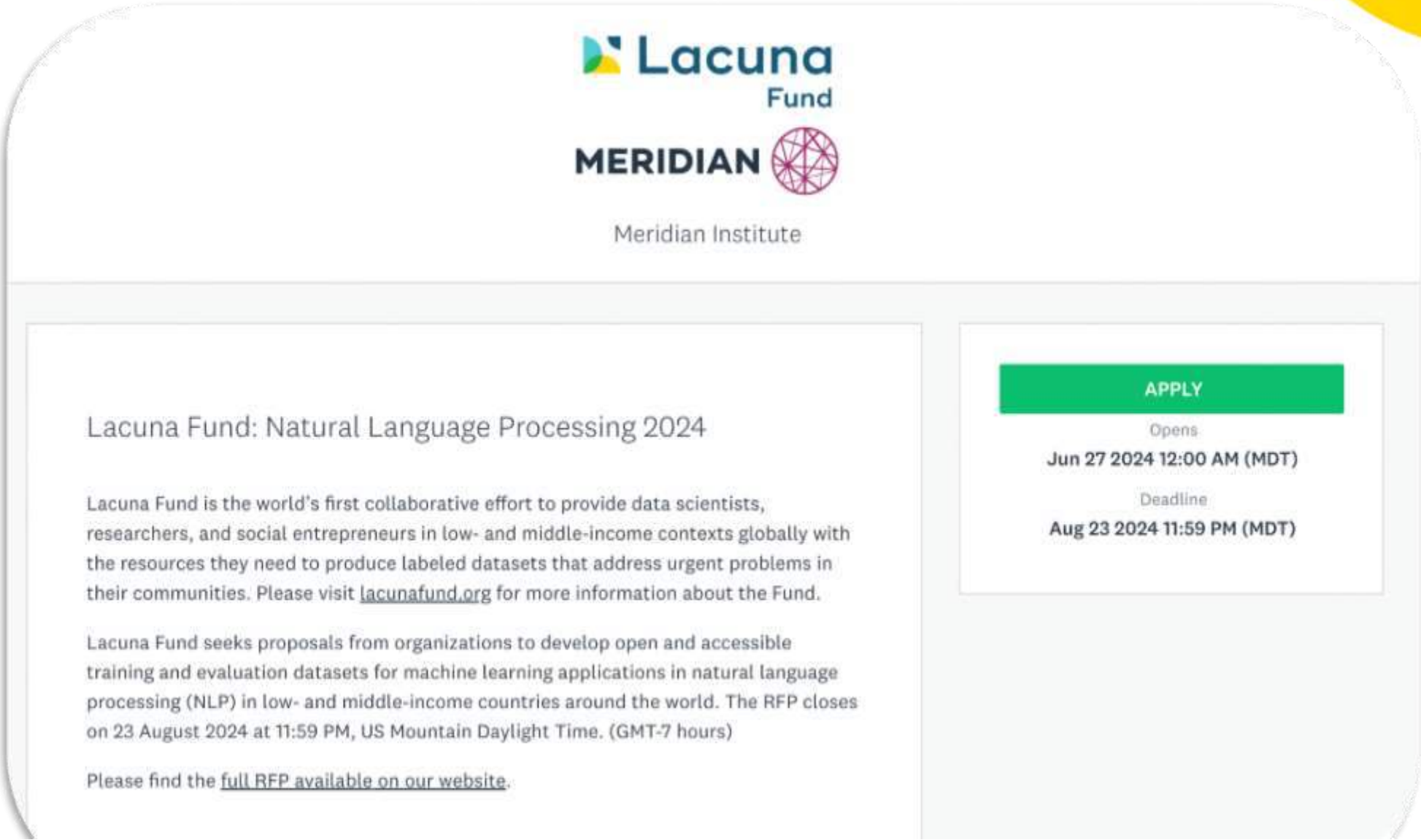
Sólo se aceptarán propuestas a través del portal de postulación SurveyMonkey

Postulación disponible en www.lacunafund.org/apply.

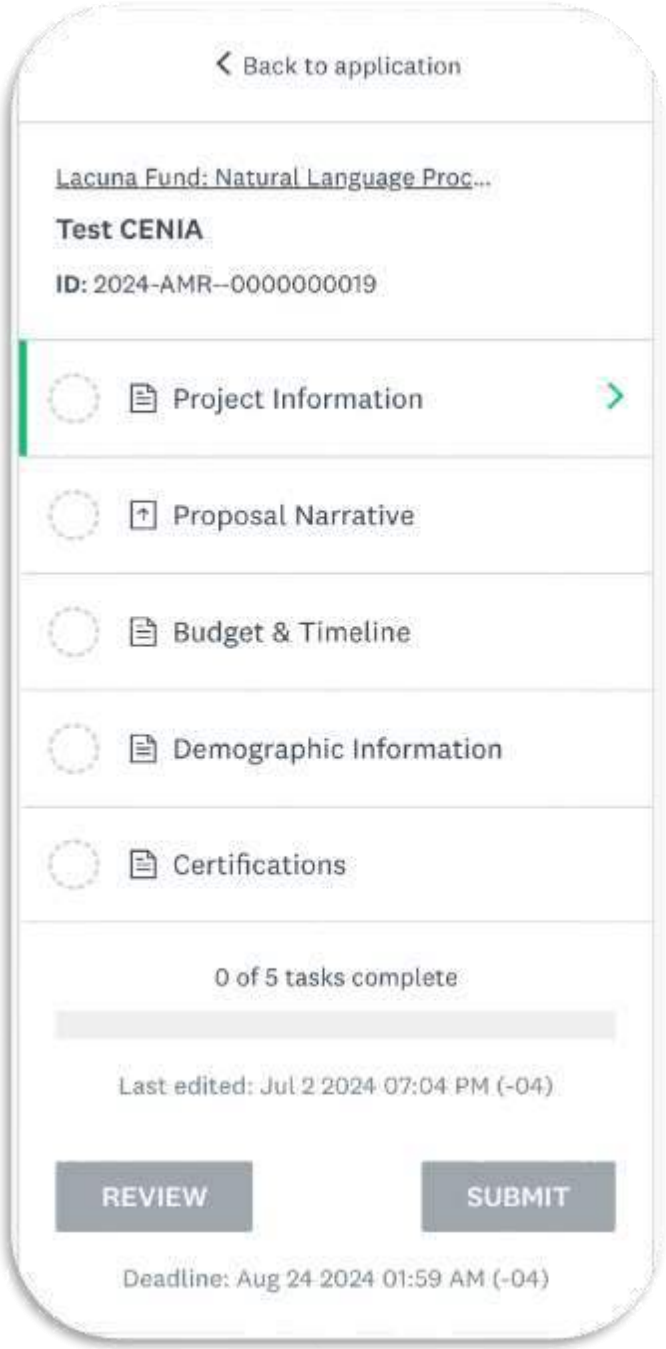
Las solicitudes pueden presentarse en **inglés**, **francés**, **portugués** y **español**.

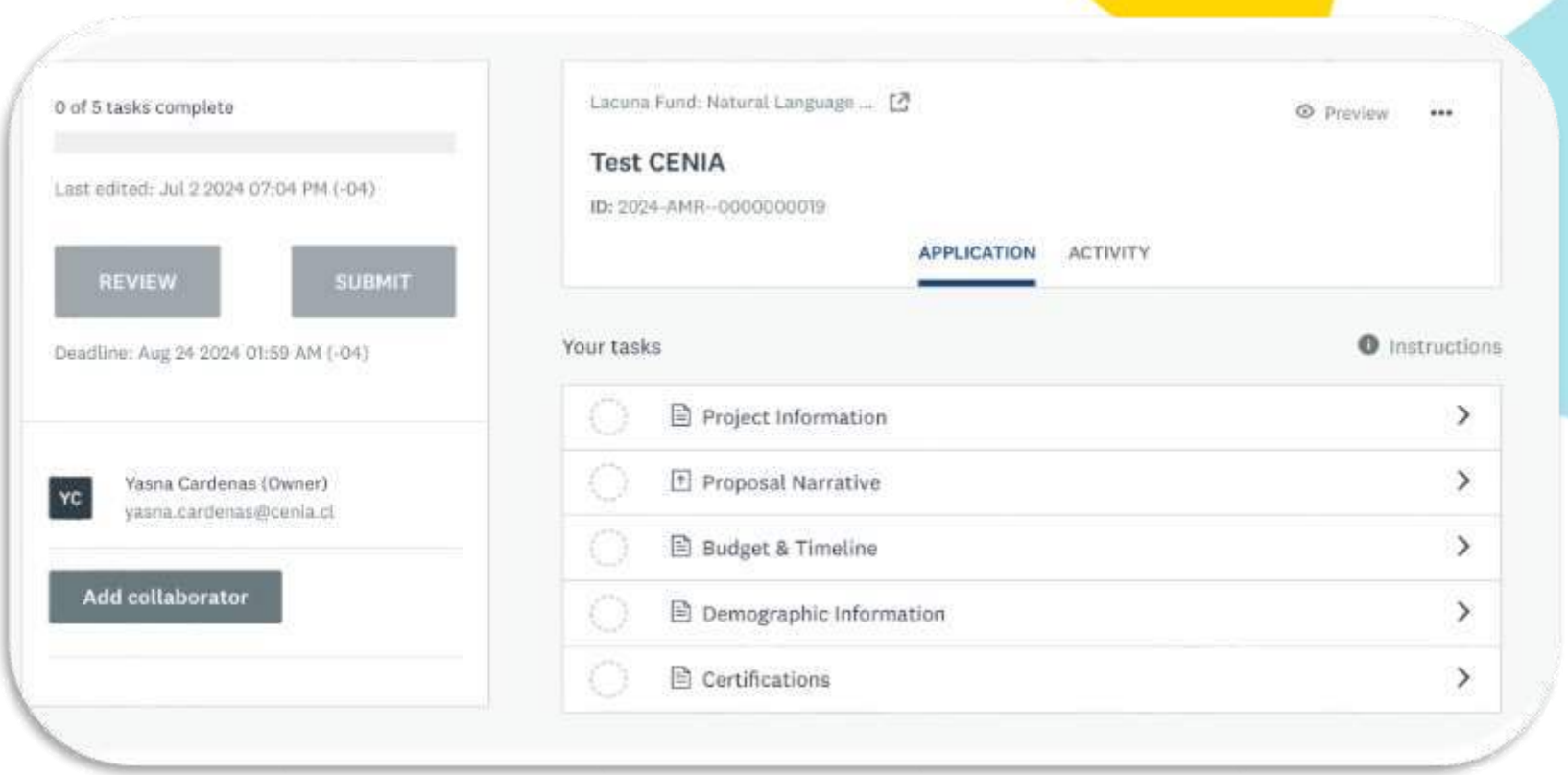
*Nota: Seleccione el idioma deseado utilizando la pestaña desplegable del portal de solicitud. Los candidatos que **presenten su solicitud en portugués podrán hacerlo a través de los portales en inglés, español o francés. Por el momento no disponemos de un portal en portugués.** Sin embargo, las propuestas presentadas en portugués en cualquier portal son aceptadas y serán revisadas.*

Postulación a través del Portal SurveyMonkey (SMA)



Postulación





¿Preguntas?

Próximos pasos

- **Envíe preguntas adicionales a secretariat@lacunafund.org antes del 12 de julio de 2024.**
- **Respuestas** publicadas en la página de solicitud del Fondo Lacuna el **29 de julio de 2024**
- **Plazo de presentación de propuestas: 23 de agosto de 2024.**