Des ensembles de données inclusifs pour l'apprentissage automatique

Traitement du langage naturel 2024
Webinaire pour les candidats

9 juillet

Lacuna Fund

Fonds Lacuna Hub

CENIA

CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL

Gestion de l'appel à propositions NLP 2024

Basé au Chili

ALVARO SOTO
CENIA

CRISTINA FLORES
CENIA

Intervenants

KATRINA GEHMAN

INSTITUT MERIDIAN, SECRÉTARIAT DU FONDS LACUNA

DR. ALBERT KAHIRA, CONSEILLER EN QUALITÉ DES DONNÉES, DATAWISE AFRIQUE

MENTORAT ET JUMELAGE, MASAKHANE

Membres du groupe consultatif technique (TAP)

AUDREY JULIA WALEGHWA MBOGHO

USIU-AFRIQUE

FELIPE DANIEL HASLER SANDOVAL, PHD

UNIVERSITÉ DU CHILI

Objectifs de la réunion

NLP 2024

- Fournir aux candidats potentiels une compréhension du Fonds Lacuna et des exigences en matière de propositions.

- Prévoir du temps pour répondre aux questions des candidats sur l'appel d'offres.

Ordre du jour

00:00	**Bienvenue, introduction,** examen de l'**ordre du jour**
00:10	**Présentation : Vue d'ensemble du Fonds Lacuna et exigences pour les appels d'offres**
00:40	**Présentation : Qualité des données et considérations relatives à l'hébergement**
00:50	**Présentation : Opportunité de mentorat et d'appariement**
01:00	**Questions et réponses**
01:25	**Prochaines étapes**
01:30	**Ajournement**

Une lacune est un vide ou une partie manquante d'un manuscrit.

Le **Fonds Lacuna** soutient la création, l'expansion et la maintenance d'ensembles de données de formation et d'évaluation équitables qui permettent aux outils d'apprentissage automatique de mieux s'attaquer aux problèmes urgents dans les contextes de revenus faibles et moyens à l'échelle mondiale.

Principes qui guident le Fonds Lacuna :

- Accessibilité
- Fonds propres
- L'éthique
- Approche participative
- Qualité
- L'impact transformationnel

Bailleurs de fonds

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

Fonds Lacuna
Structure de gouvernance

Comité de pilotage	Orientation stratégique et gouvernance globale du Fonds
Collaborateur du bailleur de fonds	Fournit au comité de pilotage des informations sur les priorités essentielles et constitue un forum pour une collaboration plus étroite avec les bailleurs de fonds.
Groupes consultatifs techniques	Étudier les processus de financement et sélectionner les propositions, fournir des informations sur les stratégies et les besoins spécifiques au domaine.
Secrétariat	Gérer les subventions et les rapports, assurer la facilitation, la communication et le soutien opérationnel.

Domaines pour les ensembles de données

Le Fonds Lacuna fournit des ressources pour la création, l'expansion ou la maintenance d'ensembles de données.

Appels à propositions en cours dans ces domaines :

- **Traitement du langage naturel (NLP)**
- **Résistance aux antimicrobiens (AMR)**

Appels à propositions antérieurs dans ces domaines :

- Agriculture
- Langue
- Santé
- Climat

Futurs appels à propositions :

- Autres domaines dans lesquels le comité de pilotage identifie un besoin

Financeur de NLP 2024

L'appel à propositions NLP 2024 est rendu possible grâce au soutien généreux de **Google.org.**

Exigences pour l'appel d'offres

NLP 2024

Appel d'offres NLP 2024 - Objet

Objectif déclaré :

L'objectif de cet appel à propositions est de soutenir les efforts visant à développer des ensembles de données ouverts et accessibles pour les applications d'apprentissage automatique liées au traitement du langage naturel (NLP) pour les **langues à faibles ressources en Amérique latine et en Afrique.**

La capacité à communiquer et à se faire comprendre dans sa propre langue est fondamentale pour l'inclusion numérique et sociétale. Les techniques de traitement du langage naturel ont permis des applications d'IA qui facilitent l'inclusion numérique et les améliorations dans les domaines de l'éducation, de la finance, des soins de santé, de l'agriculture, de la communication et des réponses aux risques naturels, entre autres. De nombreuses avancées dans le domaine du traitement du langage naturel, qu'il soit fondamental ou appliqué, ont été réalisées à partir d'ensembles de données sous licence ouverte et accessibles au public.

Toutefois, ces ensembles de données sont **rares, voire inexistants pour de nombreuses langues africaines et latino-américaines, ce qui** exclut ces populations des avantages du NLP. De nombreux modèles actuels d'apprentissage automatique sont basés sur des ensembles de données anglo-centrés et/ou traduits, manquant de nuances culturellement pertinentes et créant des modèles inutilisables pour les communautés d'Amérique latine et d'Afrique. **Lorsqu'il existe des ensembles de données pertinents, ils sont souvent basés sur des textes religieux ou judiciaires du passé, ce qui entraîne un langage obsolète et des préjugés. Il est nécessaire de disposer d'ensembles de données librement accessibles pour faciliter les technologies NLP dans les langues africaines et latino-américaines à faibles ressources** et soutenir le développement d'ensembles de données linguistiques robustes et complets qui répondent aux besoins spécifiques des communautés sous-représentées.

NLP RFP - Besoin

Le Fonds Lacuna recherche des propositions d'équipes multidisciplinaires qualifiées pour développer des ensembles de données de formation et d'évaluation ouverts et accessibles pour les applications d'apprentissage automatique pour le NLP dans les langues à faibles ressources et les cultures sous-représentées en Afrique et en Amérique latine. Les ensembles de données peuvent inclure, mais ne sont pas limités à ce qui suit :

- Ensembles de données étiquetées et non étiquetées pour les tâches NLP à faibles ressources
- Corpus de discours
- Ensembles de données pour les tâches de génération de texte
- Ensembles de données multimodales et autres données innovantes
- - Ensembles de données soutenant des tâches à forte intensité de

- Ensembles de données relatifs aux corpus de variations dialectales et aux textes et discours codés.
- Création ou augmentation d'ensembles de données textuelles et vocales spécifiques à un domaine
- Ensembles de données pour l'apprentissage automatique en linguistique
- Dans tous les ensembles de données : prise en compte de la dimension de genre et inclusion des principaux groupes

Ce que nous recherchons

- Collecte et/ou annotation de nouvelles données ;
- Annoter ou publier des données existantes ;
- Augmenter les ensembles de données existants provenant de diverses sources afin de combler les lacunes dans les données de base locales, de réduire les biais (tels que les biais géographiques, les écarts entre les sexes ou d'autres types de biais ou de discrimination), ou d'accroître la facilité d'utilisation des données et des technologies liées au NLP dans les contextes à revenus faibles et moyens ;
- Relier et harmoniser les ensembles de données existants (par exemple entre les régions, les époques, les variétés linguistiques, ainsi que les ensembles de données spécifiques à un domaine tels que les données historiques, les données sur la santé et l'éducation).

Informations sur la proposition et conditions d'éligibilité

Informations sur la proposition

- Informations sur le demandeur
- Narratif de la proposition
- Calendrier du projet et résultats attendus
- Budget
- NOTE : Opportunité de mentorat et de jumelage

Conditions d'éligibilité

Pour pouvoir bénéficier d'un financement, les organisations doivent

- être une entité à but non lucratif, une institution de recherche, une entreprise sociale à but lucratif ou une équipe de ces organisations. Les individus doivent poser leur candidature par l'intermédiaire d'un sponsor institutionnel. Les partenariats sont vivement encouragés afin de renforcer la collaboration et de maximiser les avantages découlant de l'utilisation des ensembles de données, mais seul le candidat principal recevra des fonds.
- Avoir une mission de soutien au bien sociétal, au sens large.
- Avoir son siège dans le pays ou la région où les données seront collectées. Le **présent appel se concentre géographiquement sur l'Afrique et l'Amérique latine**. Les institutions basées dans d'autres pays ou régions peuvent poser leur candidature en tant que partenaires de l'institution chef de file. Comme indiqué ci-dessus, seul le candidat principal recevra des fonds.

Conditions d'éligibilité

Pour pouvoir bénéficier d'un financement, les organisations doivent

- Disposer de toutes les autorisations nationales ou autres nécessaires pour mener à bien la recherche proposée. La procédure d'approbation peut être menée parallèlement à la demande de subvention, si nécessaire. Les frais d'approbation, le cas échéant, sont à la charge du candidat.

- Avoir la capacité technique - ou la capacité de développer cette capacité par le biais d'un partenariat décrit dans la proposition - de procéder à l'étiquetage, à la création, à l'agrégation, à l'expansion et/ou à la maintenance des ensembles de données, y compris la capacité d'appliquer les meilleures pratiques et les normes établies dans le domaine spécifique (par exemple, le traitement du langage naturel) pour permettre à de multiples entités d'effectuer des analyses d'IA/ML de haute qualité.

Dates clés

27 juin 2024 : Ouverture de l'appel à propositions

9 juillet 2024 : Webinaire pour les candidats

12 juillet 2024 : Date limite de dépôt des questions
Veuillez soumettre vos questions à secretariat@lacunafund.org

15 juillet 2024 : Date limite pour le mentorat

29 juillet 2024 : Publication des réponses

23 août 2024 : Les propositions complètes sont attendues

Printemps 2025 : Subventions accordées

Remarque : les projets proposés doivent être achevés, les ensembles de données publiés et les rapports finaux soumis d'ici **octobre 2026**. À des fins de planification, vous pouvez vous attendre à ce que les accords soient conclus et que les travaux puissent commencer d'ici **avril 2025**.

Éléments d'une proposition solide

Éléments d'une proposition solide

- **Équipe pluridisciplinaire** - expérience en science des données, traitement du langage naturel (NLP)
- **Cas d'utilisation et engagement communautaire**
- **Un énoncé** clair **du problème** et la manière dont l'ensemble de données ou l'agrégation proposés aideront à résoudre le problème.
- Spécificité concernant la taille de l'ensemble de données - la **taille et la qualité** doivent être **suffisantes** pour être utiles dans les applications futures.
- **Les partenariats** sont vivement encouragés afin de renforcer la collaboration et de maximiser les avantages découlant de l'utilisation des ensembles de données, mais seul le candidat principal recevra des fonds.

Éléments d'une proposition solide

- Considérations relatives à **l'équité** (sexe, statut socio-économique, ethnicité, etc.)
- Prise en compte et planification des problèmes potentiels liés à la **protection de la vie privée et à l'éthique**
- Planifier la **gestion des données et l'octroi de licences**
- Travailler dans plusieurs domaines lorsque c'est possible et pertinent (régions, temps, variétés linguistiques).
- **Le budget** est adapté à la taille de l'ensemble de données produit

Engagement communautaire

- Décrire les consultations antérieures et/ou la collaboration proposée avec les bénéficiaires prévus.
  - Quand et où vous avez rencontré / rencontrerez les partenaires
  - Comment les partenaires seront-ils impliqués ?
    - Identifier les besoins en données
    - Collecte de données, étiquetage
    - Gouvernance des données
    - Utilisation des données
  - Comment les partenaires bénéficieront-ils de l'ensemble de données nouveau/élargi ?

Vie privée et éthique

- Expliquez comment votre équipe abordera la question :

- a) le respect de la vie privée,
- b) le risque d'utilisation abusive en aval,
- c) les éventuels vecteurs de discrimination (par exemple, le sexe), et
- d) des conditions de travail justes et équitables, si des étiqueteurs rémunérés participent au projet.

- Décrivez le processus que vous utiliserez pour détecter les problèmes éthiques potentiels (par exemple, un comité d'examen institutionnel, etc.)

Plan de développement durable et de communication

- Décrivez comment le jeu de données sera maintenu et/ou étendu au-delà du financement initial (par exemple, par le biais d'un modèle de référence, d'applications ML résultantes, par une communauté dédiée ou un groupe de parties intéressées avec un modèle de gouvernance solide pour le jeu de données ouvert) et comment un cas d'utilisation potentiel pourrait soutenir le projet.

- Décrivez les activités de communication visant à faire connaître le(s) jeu(x) de données. Il peut s'agir d'activités de mise en réseau avec des utilisateurs potentiels de données, de la présentation de l'ensemble de données lors d'une conférence, de l'organisation d'un atelier sur la durabilité de l'ensemble de données avec les parties prenantes intéressées ou de la mise en place d'un comité de durabilité.

Normes et partage des données

- Facile à trouver - facile à trouver sur une plateforme publique et largement utilisée
- Accessible - format ouvert (CCBY 4.0 or CCBY SA 4.0)
- Interopérabilité - format des données
- Réutilisable - métadonnées

- Durable - plan de maintenance
- Partagé - communauté d'utilisateurs engagée et plan de partage de l'ensemble des données une fois terminé

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

Lacuna Fund Intellectual Property Policy: **IP-Policy_LacunaFund.pdf**

Budget et coûts autorisés

- Fournissez un budget pour la réalisation de l'ensemble de données proposé via le portail SurveyMonkey Apply. Ce budget doit être formaté selon le modèle de budget du Fonds Lacuna, disponible sur le portail de candidature.

Le montant total disponible est d'environ 1 million de dollars. Nous souhaitons financer des projets dans chacune des régions cibles (Afrique, Amérique latine) et prévoyons de soutenir 6 à 8 petits projets avec des budgets allant jusqu'à 100 000 USD et 2 à 3 projets plus importants et plus complexes avec des budgets allant de 100 à 250 000 USD. Le comité consultatif technique évaluera la faisabilité et la pertinence du budget ainsi que le lien entre le budget et la description de la subvention dans le cadre des critères de sélection.

Budget et coûts autorisés

Les budgets peuvent inclure, sans s'y limiter, les coûts suivants

- Renforcement des capacités en matière de collecte de données et d'assurance/contrôle de la qualité ;
- Collecte de données (y compris une compensation équitable pour la fourniture de données) ;
- l'étiquetage des données (y compris une compensation équitable pour l'étiquetage des données) ;
- AQ/CQ ou vérification ;
- Post-traitement des données ;
- Publication des données ;
- Modèle de base
- Octroi de licences.
- Publication des résultats en libre accès.
- Il est temps de préparer une déclaration de données pour l'ensemble de données.
- Les efforts de crowdsourcing, tels que les "label-a-thons".
- Stockage des données.
- Puissance de calcul.
- Atelier
- Activités de communication, y compris la participation à deux conférences au maximum pour présenter les ensembles de données.

Considérations sur la qualité des données

Hébergement et accessibilité - Principes de Lacuna

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

Hébergement - Lignes directrices de Lacuna

- Attribue un identifiant d'objet numérique (DOI) aux ensembles de données ou permet d'en joindre un dans les métadonnées.

- est indexé par les principaux moteurs de recherche (par exemple, Google Dataset Search ou des outils similaires).

- Il/elle est fiable et persévérant(e)

Très utile :

- Quantifie le nombre de consultations et de téléchargements de la page d'atterrissage pour l'ensemble de données.

- Collecte des informations de contact pour les téléchargements d'ensembles de données de manière à maximiser la conversion.

Hébergement de l'ensemble de données

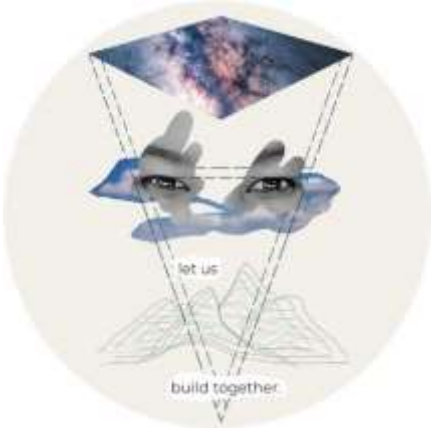
La documentation et l'hébergement proposés sont conformes aux lignes directrices du Lacuna Fund en matière d'hébergement et de documentation des ensembles de données.

Le Fonds Lacuna demande aux bénéficiaires d'inclure la documentation suivante lorsqu'ils soumettent des ensembles de données :

- Fichier de métadonnées
- Fiche technique
- Identifiant d'objet numérique (DOI)

Opportunité de mentorat : Le programme de mentorat Masakhane Opportunité :

Fondation de recherche Masakhane



Le programme de mentorat de Masakhane

- **Mission :** La MRF est une organisation de base dont la mission est de renforcer et de stimuler la recherche en NLP dans les langues africaines, pour les Africains, par les Africains. L'objectif de MRF est de permettre aux Africains de façonner et de s'approprier ces avancées technologiques en faveur de la dignité humaine, du bien-être et de l'équité, par le biais d'une construction communautaire inclusive, d'une recherche participative ouverte et de la multidisciplinarité.

- La FRM offrira aux candidats d'Afrique et d'Amérique latine une occasion spéciale de rejoindre la communauté de la FRM et de mettre en relation ceux qui sont intéressés avec un mentor qui examinera votre projet de proposition et discutera des moyens de le renforcer.

Le programme de mentorat de Masakhane

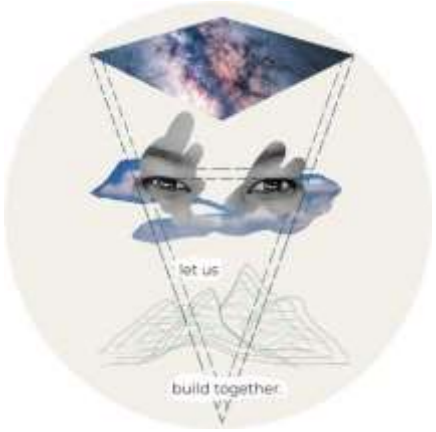
- Postulez sur le **formulaire Google** ci-dessous https://forms.gle/8u5GobjKXYo7gujt5
- Fournir une brève description (résumé de 250 mots) de la proposition d'ensemble de données.
- Les mentorés peuvent demander différentes formes d'assistance telles que
    - "discuter des lacunes dans le domaine du NLP à faibles ressources",
    - "Rédaction d'une proposition de recherche", et
    - "préparer les budgets".

Le programme de mentorat de Masakhane

- Les candidats intéressés sont encouragés à postuler pour une session de mentorat au moins 6 semaines avant la date limite de dépôt des propositions, soit le **15 juillet 2024.**

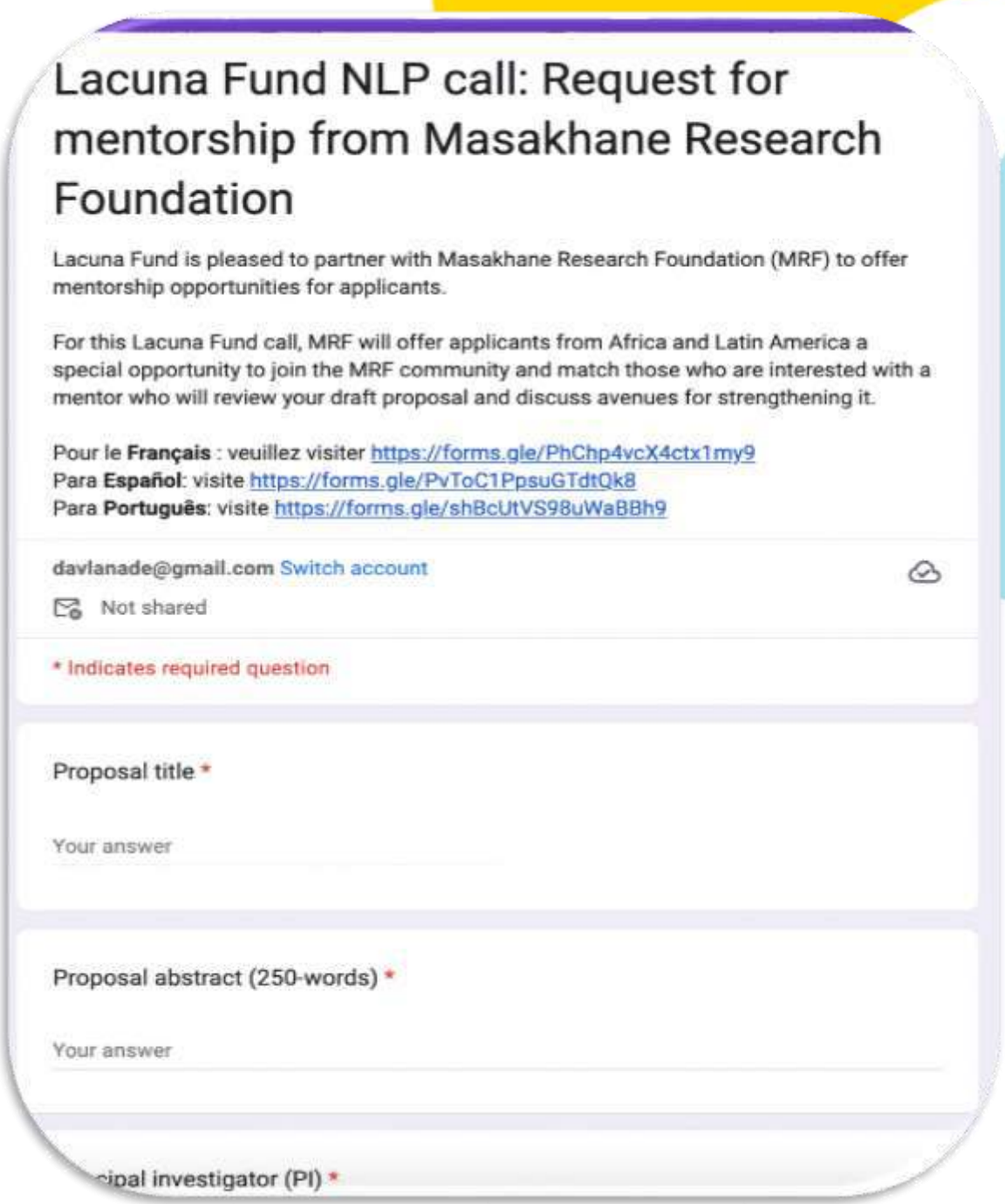
- Les mentors seront attribués selon le principe du premier arrivé, premier servi.

- Tous les candidats sont tenus de lire et de respecter le **code d'éthique et de conduite** du programme de mentorat.



Pour plus d'informations, veuillez consulter le site [: https://lacunafund.org/apply/](https://lacunafund.org/apply/)

Des questions ?
Courriel :
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

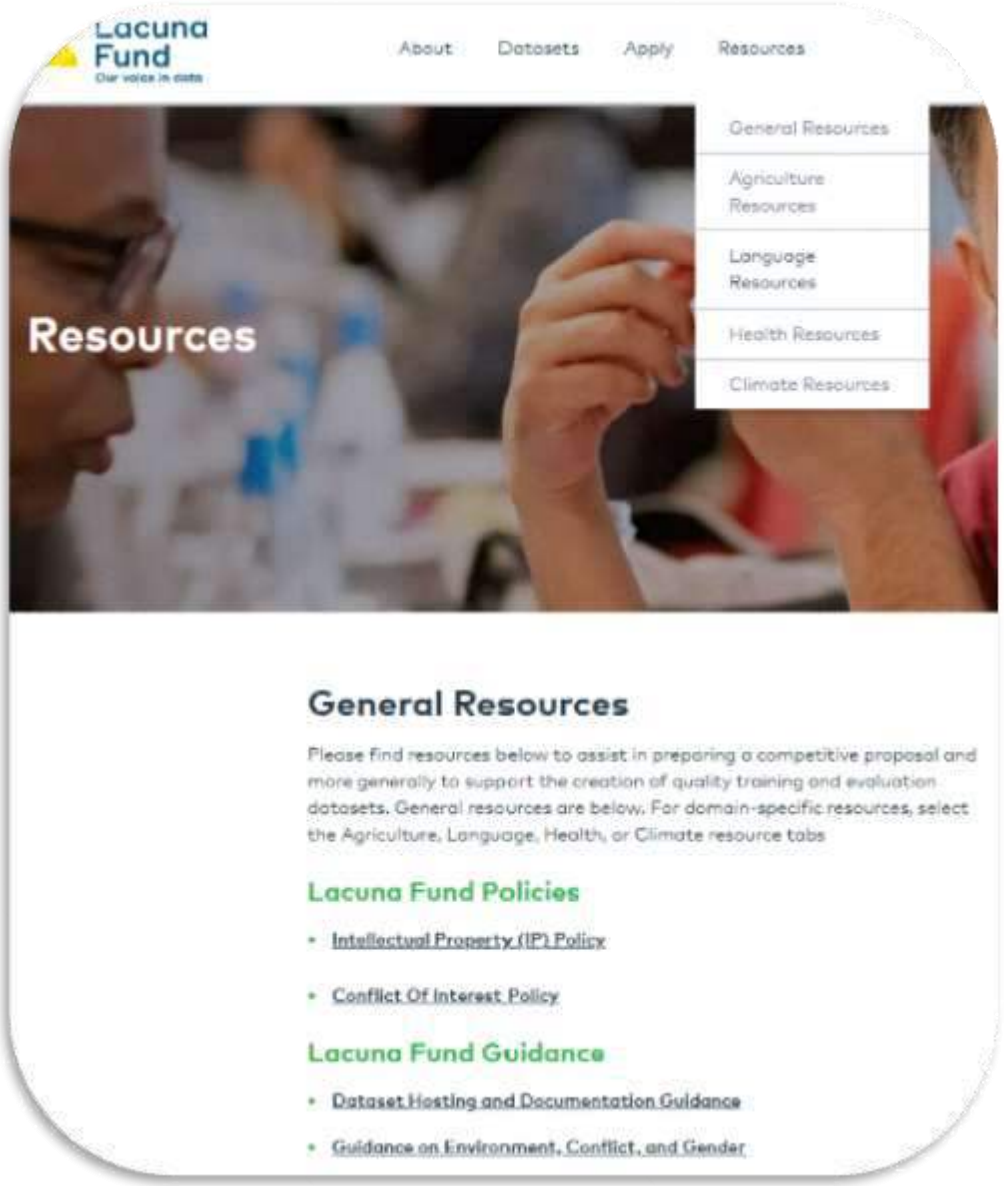
Proposal abstract (250-words) *

Your answer

cipal investigator (PI) *

Ressources

Ressources



Le site web du Fonds Lacuna comprend diverses ressources, telles que des références pertinentes sur la **qualité des données et la documentation**, afin d'aider les candidats à préparer une candidature compétitive.

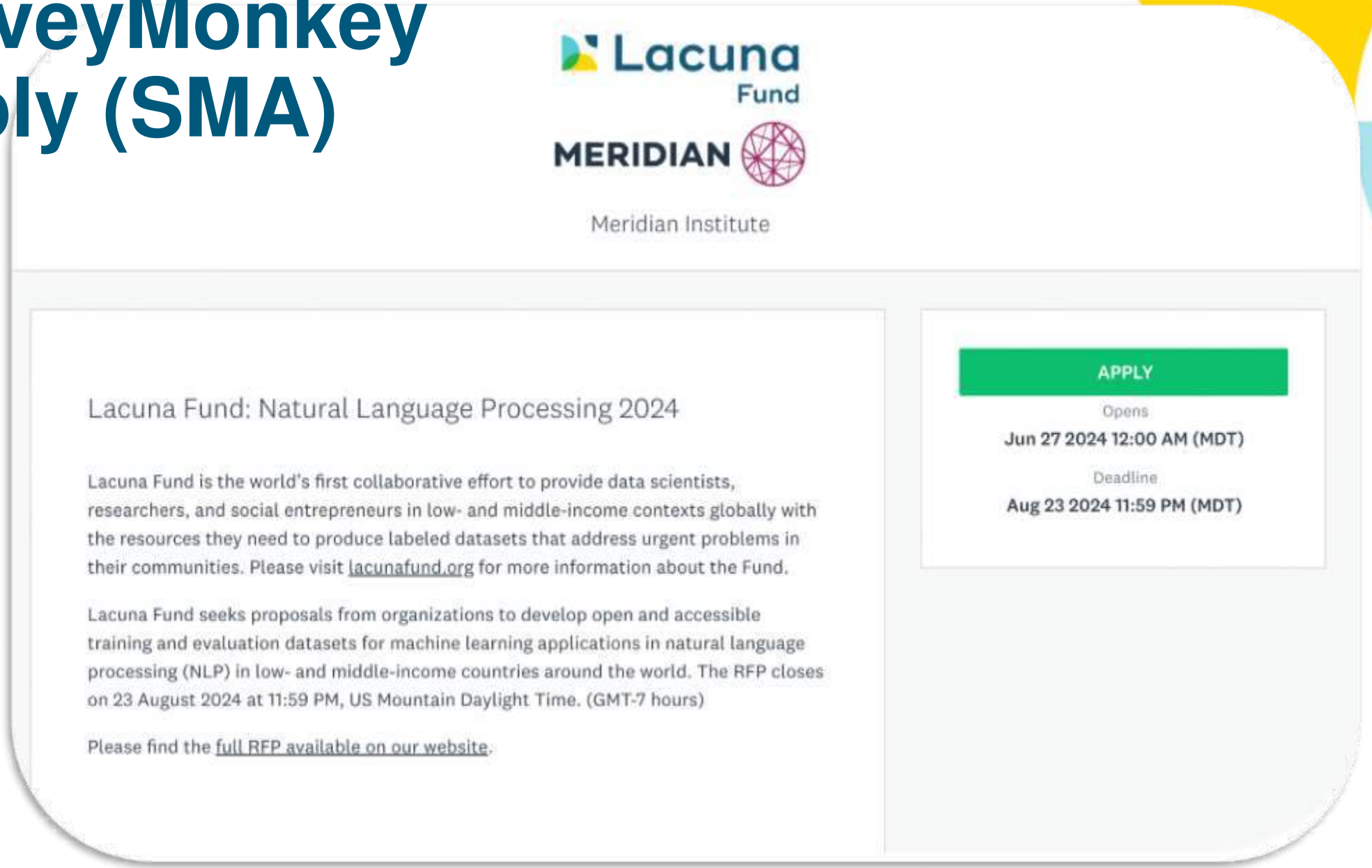
Soumission de la proposition

Les propositions ne seront acceptées que par l'intermédiaire du portail de candidature SurveyMonkey Apply, disponible à l'adresse www.lacunafund.org/apply.

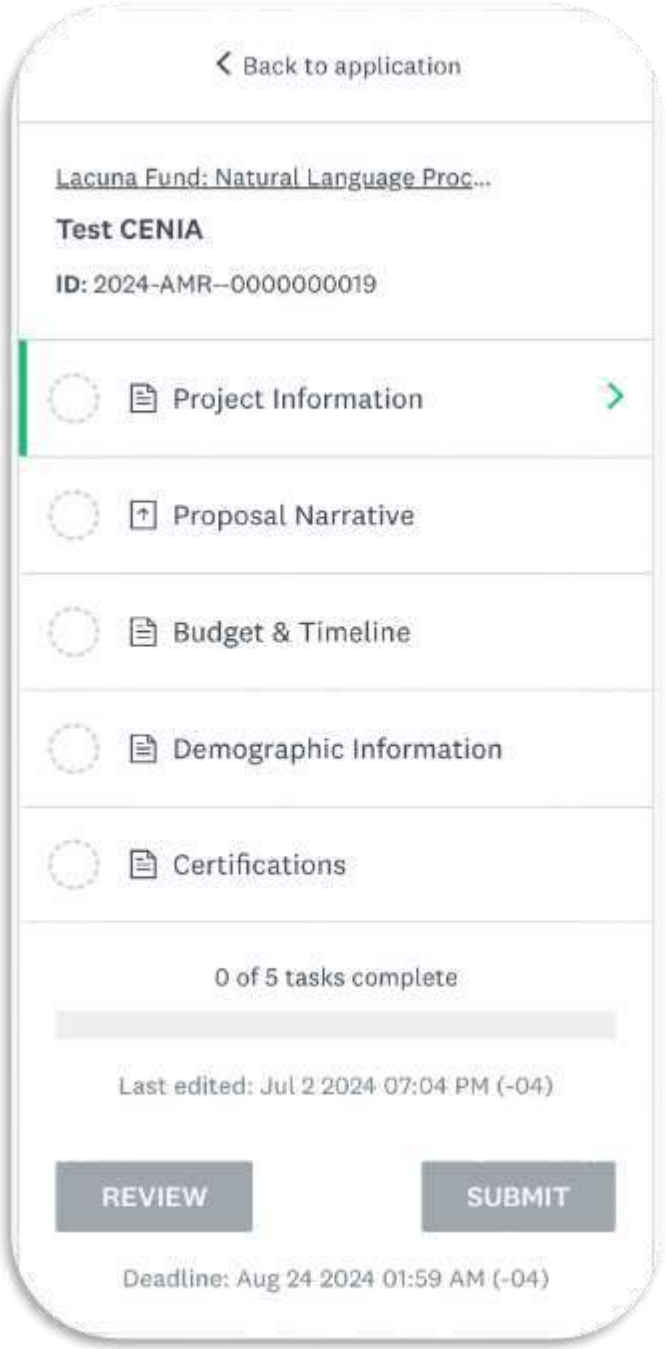
Les candidatures peuvent être soumises en **anglais**, **français**, **portugais** et **espagnol**.

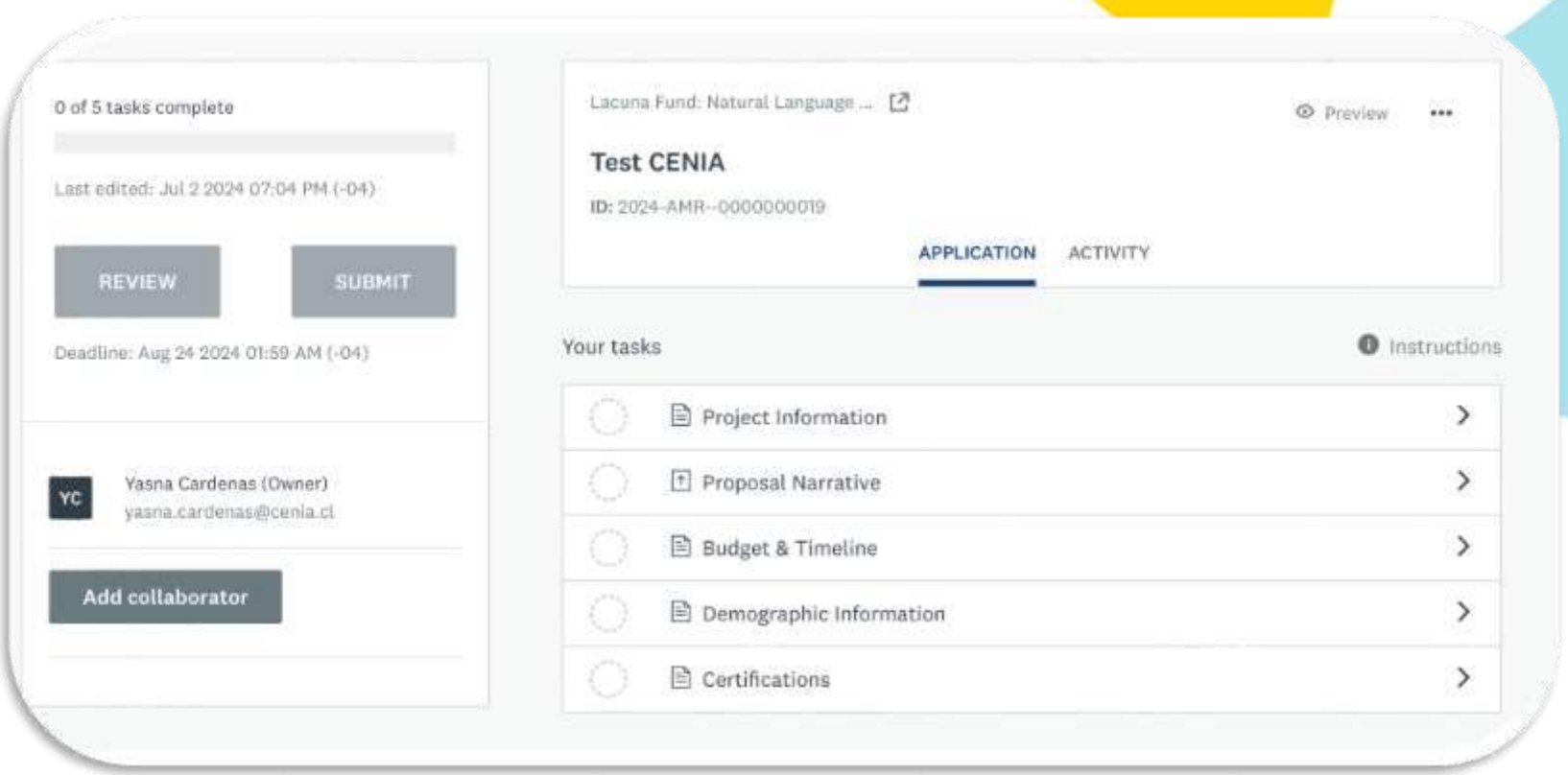
*Remarque : sélectionnez la langue de votre choix à l'aide de l'onglet déroulant du portail de candidature. Les **personnes qui soumettent leur candidature en portugais peuvent utiliser les portails en anglais, en espagnol ou en français. Le portail portugais n'est pas encore disponible.** Toutefois, les propositions soumises en portugais dans n'importe quel portail sont acceptées et seront examinées.*

Portail SurveyMonkey Apply (SMA)



Application





Des questions ?

Prochaines étapes

- **Soumettre des questions supplémentaires à secretariat@lacunafund.org avant le 12 juillet 2024.**
- **Réponses affichées** publiquement sur la page Apply du Fonds Lacuna le **29 juillet 2024**
- **Les propositions sont attendues pour le 23 août 2024.**