# Appel à propositions : Traitement automatique du langage naturel (TALN) 2024

## Lacuna Fund : Our Voice in Data

**27 juin 2024**

# Table des matières

1 – INTRODUCTION ..... 3

VUE D'ENSEMBLE ET OBJECTIFS DU LACUNA FUND ..... 3
PHILOSOPHIE EN MATIERE D'OCTROI DE SUBVENTIONS ..... 3

2 – VUE D'ENSEMBLE ..... 4

CRITERES D'ELIGIBILITE APPLICABLES AUX ORGANISATIONS ..... 4
PROCESSUS DE SELECTION ET CRITERES D'EVALUATION ..... 4
CALENDRIER ..... 6

3 – OBJET ET BESOINS ..... 7

OBJET ..... 7

4 – INFORMATIONS SUR LES PROPOSITIONS ..... 9

INFORMATIONS SUR LE CANDIDAT ..... 9
EXPOSE DE LA PROPOSITION ..... 10
CALENDRIER DU PROJET ET LIVRABLES ..... 13
BUDGET ..... 13

# 1 – Introduction

## Vue d'ensemble et objectifs du Lacuna Fund

Lacuna Fund soutient la création, le développement et la tenue à jour d'ensembles de données permettant une application fiable et équitable des outils d'apprentissage machine (AM) présentant une valeur sociale élevée dans les contextes de faibles et moyens revenus au niveau mondial.

Le Fonds poursuit plusieurs objectifs :

- Verser des fonds aux institutions en vue de créer, de développer et/ou de tenir à jour des ensembles de données qui comblent les lacunes et réduisent les biais dans les données étiquetées utilisées pour la formation et/ou l'évaluation des modèles d'apprentissage machine.
- Permettre aux populations mal desservies de profiter des avancées offertes par l'IA.
- Favoriser une meilleure compréhension par la communauté de l'apprentissage machine et les organisations philanthropiques de la manière de financer le plus efficacement et économiquement possible l'élaboration et la tenue à jour d'ensembles de données étiquetées dans le respect des principes d'équité.

## Philosophie en matière d'octroi de subventions

Lacuna Fund privilégie une approche collaborative et pilotée localement pour la création, le développement et la tenue à jour d'ensembles de données. Nous considérons que l'utilité et l'actualisation pérennes des données ouvertes imposent de s'appuyer sur une communauté investie et concernée par ces données. Nous considérons également la collaboration comme étant un moyen d'améliorer la qualité et l'impact des ensembles de données et de promouvoir une culture de la coopération dans le domaine.

Lacuna Fund espère financer des ensembles de données qui contribuent aux multiples applications à forte valeur sociale, que ce soit au travers de la recherche, de l'innovation commerciale ou de l'amélioration des services du secteur public. **Bien que la section 3 « Objet et besoins » expose les besoins définis par le Groupe consultatif technique (Technical Advisory Panel, TAP), Lacuna Fund accueille toutes les idées novatrices dans ce domaine mettant en évidence un avantage clairement articulé aligné sur les critères de sélection présentés ci-après.**

Cet appel à propositions est soutenu par Google.org.

# 2 – Vue d’ensemble

## Critères d’éligibilité applicables aux organisations

Lacuna Fund entend rendre son financement accessible à un maximum d’organisations dans l’univers de l’IA pour le bien social et cultiver les capacités et les organisations émergentes dans ce domaine.

**Pour pouvoir bénéficier d’un financement, les organisations doivent remplir les critères suivants :**

- être une entité à but non lucratif, un institut de recherche, une entreprise sociale à but lucratif ou une équipe composée de ce type d’organisations. Pour présenter leur projet, les particuliers doivent recourir à un promoteur institutionnel. Les partenariats sont fortement 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 du bien sociétal, au sens large ;
- avoir leur siège social dans le pays ou la région où les données seront collectées. L'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 ;
- disposer de toutes les autorisations requises, nationales ou autres, pour poursuivre les recherches proposées. Le cas échéant, le processus d’autorisation pourra être mené en parallèle de la demande de subvention. Les éventuels frais d’autorisation sont à la charge du candidat ;
- disposer de la capacité technique – ou être en mesure de développer cette capacité grâce à un partenariat décrit dans l’appel à propositions – pour procéder à l’étiquetage, à la création, au développement et/ou à la tenue à jour des ensembles de données, y compris la capacité à mettre en œuvre les bonnes pratiques et les normes établies dans le domaine visé (par ex. Traitement automatique du langage naturel, TALN) afin de permettre à plusieurs entités de réaliser des travaux analytiques de haute qualité pour l’IA/l’AM.

## Processus de sélection et critères d’évaluation

Lacuna Fund est à la recherche de propositions visant à créer, développer, consolider et/ou débloquer des ensembles de données étiquetées pour les applications d’apprentissage machine qui permettront d’obtenir des solutions en matière de Traitement automatique du langage naturel (TALN) équitables en Afrique et en Amérique latine. Lacuna Fund et ses partenaires procèdent à une sélection initiale des propositions pour vérifier l’éligibilité organisationnelle et la faisabilité. À l’issue de cette sélection initiale, un Groupe consultatif technique composé d’experts, d’utilisateurs de données et de parties prenantes pour le domaine visé évalue les propositions sur la base des critères de sélection énoncés ci-après. Les membres du Groupe consultatif technique ne peuvent pas soumettre une proposition pour répondre à un appel à propositions pour lequel ils sont évaluateurs (voir la Politique du Lacuna Fund en matière de conflit d’intérêts).

**Le Groupe consultatif technique pour cet appel examinera les soumissions et sélectionnera une série de propositions pour financement. Les propositions seront sélectionnées sur la base du niveau de satisfaction des critères suivants, qui sont basés sur les principes qui guident les opérations du Lacuna Fund :**

- **Qualité** – l'organisation ou l'équipe proposant le projet dispose d'experts qualifiés dans : a) le domaine visé, b) l'apprentissage machine, et c) la gestion de données. Les collaborations avec les agences gouvernementales et les groupes communautaires sont encouragées. L'équipe fournit des cas d'utilisation clairs pour l'ensemble de données. Le porteur de projet situe l'ensemble de données proposé par rapport aux ressources existantes (ou au manque de ressources) dans le domaine et propose l'utilisation d'outils et techniques efficaces de collecte et d'étiquetage de données permettant d'accélérer la collecte, le nettoyage et le partage des données.
- **Impact transformationnel** – le projet rend le ou les ensembles de données existants plus représentatifs, inclusifs et/ou durables ou crée un nouvel ensemble de données étiquetées de grande valeur à destination d'une population mal desservie ou pour apporter une réponse à un problème. Une proposition peut être considérée comme transformationnelle si elle a le potentiel de s'attaquer à un problème particulièrement important ou opportun en matière d'équité dans le domaine de l'AM/IA ou si elle a une portée significative en termes de nombre de personnes ou de régions géographiques mal desservies.
- **Équité** – l'équipe énonce le problème d'équité qu'elle propose de traiter et décrit comment l'ensemble de données comblera les lacunes et rendra le domaine plus représentatif et équitable. La proposition est assortie d'un argumentaire convaincant démontrant comment l'ensemble de données permettra aux populations vulnérables et mal desservies d'accéder plus largement aux avantages de l'AM/IA.
- **Approche participative** – l'équipe est implantée dans la région géographique où les données seront collectées pour garantir une tenue à jour et un usage pérennes de l'ensemble de données par la communauté locale. Les partenaires nationaux sont impliqués dans les éléments stratégiques du projet (au-delà de la collecte de données). La proposition décrit comment l'équipe fera participer les parties prenantes concernées, obtiendra le consentement éclairé pour la collecte et l'utilisation des données, et partagera les résultats du projet avec les fournisseurs de données et/ou la communauté.
- **Éthique** – le projet propose un plan pour aborder et peut passer avec succès un examen sur les questions d'éthique (par exemple, par un comité d'examen institutionnel) portant sur les aspects suivants : a) les questions liées au respect de la vie privée ; b) le risque d'utilisation abusive de l'ensemble de données en aval ; c) les vecteurs de discrimination potentiels (par exemple, la discrimination fondée sur le genre) ; et d) les conditions de travail justes et équitables si des étiqueteurs rémunérés participent au projet. Les objectifs et la méthodologie proposés par le projet sont impartiaux et éthiques.
- **Durabilité et communications** – le projet comporte un plan visant à garantir la durabilité et la tenue à jour future de l'ensemble de données, par exemple grâce à un modèle de référence, aux applications d'AM issues du projet, par une communauté spécifique ou un groupe de parties intéressées (organisations à but lucratif ou non lucratif), un modèle de gouvernance robuste pour l'ensemble de données ouvert, et les cas d'utilisation potentiels de l'ensemble de données. Le plan peut indiquer qui mettra à jour et gérera l'ensemble de données, les sources de financement potentielles ; les stratégies de mobilisation proposées

pour les populations touchées et les utilisateurs de données ; les projets de présentation des ensembles de données lors d'une ou plusieurs conférences ; l'organisation d'un atelier sur la pérennité des ensembles de données avec les parties prenantes intéressées ; ou la mise en place d'un comité de pérennité, et les mesures visant à maintenir les données ouvertes et accessibles.

- **Faisabilité** – le projet est réalisable en fonction du budget et de l'étendue des travaux proposés.
- **Accessibilité** – l'ensemble de données sera largement accessible dans le cadre d'une licence ouverte ou, en cas d'impossibilité, un argumentaire convaincant expose les raisons d'un système d'octroi de licence plus strict dans le but de protéger la vie privée ou de prévenir tout préjudice. Les sous-bénéficiaires donneront la priorité à la diffusion de la propriété intellectuelle sous une structure de licence de source ouverte permissive, telle qu'Apache 2.0 pour tout code ou autres inventions, ou CC-BY 4.0 International, ou CC BY-SA 4.0 pour toute autre propriété intellectuelle (par exemple, des œuvres créatives qui ne sont pas du code ou qui ne sont pas brevetables). La documentation proposée et l'hébergement sont alignés sur les orientations décrites dans le document du Lacuna Fund intitulé Dataset Hosting and Documentation Guidance (Orientations en matière d'hébergement des données et de documentation).

# Calendrier

| | |
|---|---|
| L'appel à propositions est publié le | 27 juin 2024 |
| Webinaire pour les candidats | 9 juillet 2024 |
| Date limite pour les questions/réponses<br><br>Veuillez adresser vos questions à secretariat@lacunafund.org | 12 juillet 2024 |
| Date limite pour le mentorat | 15 juillet 2024 |
| Publication des réponses | 29 juillet 2024 |
| Remise des propositions complètes | 23 août 2024 |

**Période de questions/réponses :** Toutes les questions concernant l'appel à propositions doivent être adressées par courrier électronique à secretariat@lacunafund.org en précisant « NLP 2024 RFP Question (Question Appel à propositions TALN 2024)» dans l'objet. Les questions adressées d'ici le 12 juillet seront anonymisées et leur réponse sera publiée le 29 juillet dans un document posté sur la page « Soumettre un projet » du site web du Lacuna Fund et envoyées à tous les candidats par email.

# 3 – Objet et besoins

## Objet

L'objet 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 machine liées au traitement automatique du langage naturel (TALN) pour les langues et cultures à faibles ressources en Afrique et en Amérique latine.

La capacité à communiquer et à être compris dans sa propre langue et dans son propre contexte culturel est indispensable à l'inclusion numérique et sociétale. Les techniques de traitement automatique du langage naturel ont le potentiel de permettre des applications d'IA qui facilitent l'inclusion numérique et des améliorations dans l'éducation, la finance, les soins de santé, l'agriculture, la communication et les réponses aux risques naturels, entre autres. De nombreuses avancées dans le domaine du TALN, tant fondamental qu'appliqué, ont été réalisées grâce à des ensembles de données sous licence ouverte et accessibles au public.

Cependant, ces ensembles de données sont rares, voire inexistants pour de nombreuses langues africaines et latino-américaines, ce qui exclut ces populations des avantages du TALN. De nombreux modèles actuels d'apprentissage machine (AM) sont basés sur des ensembles de données anglo-centrés et/ou traduits, manquant de nuances culturellement pertinentes et créant des modèles biaisés ou inutilisables pour les communautés d'Afrique et d'Amérique latine. Lorsque des ensembles de données pertinents existent, ils s'appuient souvent sur des textes religieux ou judiciaires du passé, produisant un langage désuet et des biais. Il est nécessaire de disposer d'ensembles de données librement accessibles afin de mettre les technologies du TALN au service des langues à faibles ressources en Afrique et en Amérique latine, et de soutenir le développement d'ensembles de données linguistiques solides et culturellement appropriés qui répondent aux besoins spécifiques des communautés sous-représentées.

## Besoins

Lacuna Fund est à la recherche des propositions d'équipes qualifiées et pluridisciplinaires désireuses d'élaborer des ensembles de données de formation et d'évaluation ouverts et accessibles pour des applications d'apprentissage machine pour le TALN dans des langues à faibles ressources et les cultures sous-représentées en Afrique et en Amérique latine.

Les propositions peuvent inclure, mais ne sont pas limitées à :

- la collecte et/ou l'annotation de nouvelles données ;
- l'annotation ou la publication de données existantes ;
- l'augmentation des ensembles de données existants provenant de diverses sources pour combler les lacunes de données de terrain, réduire les biais (tels que les biais géographiques ou d'autres types de biais ou de discrimination) ou augmenter l'utilisabilité des données et des technologies liées au TALN dans les contextes de revenus faibles et intermédiaires ;
- la mise en relation et l'harmonisation des 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).

Bien que Lacuna Fund se concentre principalement sur la création, l'expansion et la tenue à jour d'ensembles de données, les propositions peuvent inclure le développement d'un ou de plusieurs modèles de référence pour assurer la qualité de l'ensemble de données financé et/ou pour faciliter l'utilisation d'ensembles de données dans des applications socialement bénéfiques.

Le Groupe consultatif technique estime qu'il est nécessaire de disposer d'ensembles de données de formation et d'évaluation qui tiennent compte de la diversité linguistique et des nuances culturelles de l'Afrique et de l'Amérique latine. Il s'agit notamment d'ensembles de données sur l'argot régional, les expressions idiomatiques, les variétés linguistiques locales ou les dialectes, ainsi que des données pertinentes d'un point de vue culturel. Ces ensembles de données sont essentiels pour développer des outils de traitement automatique du langage naturel plus inclusifs et plus efficaces, capables de répondre aux besoins spécifiques des communautés linguistiques culturellement diversifiées.

Nous recherchons des ensembles de données identifiés par des experts locaux et conçus pour répondre à des besoins définis localement. Les exemples suivants ne sont fournis qu'à titre d'illustration.

**Les ensembles de données peuvent inclure, sans s'y limiter :**

- les **ensembles de données étiquetées et non étiquetées pour les tâches de TALN dans les langues à faibles ressources**, soutenant le développement de modèles d'apprentissage machine précis et efficaces. Les tâches en aval des ensembles de données étiquetées peuvent inclure, sans s'y limiter : la réponse aux questions et l'IA conversationnelle, les ensembles de données d'analyse des sentiments, la détection des préjugés sociaux, la détection des discours haineux et la contre-parole, la détection des fausses informations et de la désinformation ; le résumé automatique de texte ou d'autres tâches de compréhension et de génération de langage naturel, ou des ressources pour soutenir l'enseignement du TALN en collaboration avec les communautés. Les ensembles de données non étiquetées comprennent des corpus de textes qui peuvent être utilisés pour la formation et l'évaluation de modèles vocaux ;
- des **corpus vocaux**, y compris les ensembles de données permettant la reconnaissance automatique de la parole (RAP) qui permet aux analphabètes ou aux groupes de personnes défavorisées d'accéder à des informations et/ou à des services dans des langues à faibles ressources ;
- les **ensembles de données pour les tâches de génération de texte**, en particulier autres que la traduction automatique ;
- les **ensembles de données multimodales et autres données innovantes**, telles que le sous-titrage vidéo ou audio, les questions-réponses visuelles ou d'autres interactions image-texte ;
- les **ensembles de données qui soutiennent les tâches à forte intensité de connaissances**, telles que l'assurance qualité et la génération augmentée de récupération (RAG);
- les **ensembles de données relatifs aux corpus de variations dialectales et à l'alternance codique dans les textes et le discours**, y compris la saisie des variations linguistiques (argot régional, expressions idiomatiques, données culturellement pertinentes) dans les langues à faibles ressources riches en dialectes et dans les communautés linguistiques où l'alternance codique est courante ;

- la **création ou l'augmentation d'ensembles de données textuelles et vocales spécifiques à un domaine**, comme les soins de santé, les noms de lieux, l'agriculture ou l'éducation, qui permettent des applications ayant un impact social important ; l'exploration des cadres d'augmentation générative des données pour inclure un vocabulaire, une sémantique, une morphologie et une syntaxe spécialisés dans un domaine ;
- les **ensembles de données soutenant l'apprentissage machine pour la linguistique**, pour la préservation et la revitalisation des cultures marginalisées et des aspects des langues sous-représentées que ces cultures considèrent comme importants pour leur santé, leur dignité, leur environnement et leur bien-être. Ces ensembles de données peuvent inclure des annotations phonétiques, morphologiques et syntaxiques, ainsi que des outils automatisés pour exécuter ces tâches si le groupe social concerné le souhaite ;
- **dans tous les ensembles de données : prise en compte de la dimension de genre et inclusion des principaux groupes vulnérables**, y compris l'atténuation des biais pour les personnes vivant dans des contextes humanitaires et de conflit, ainsi que pour celles qui se trouvent à l'intersection de plusieurs groupes socio-économiques (par exemple, handicap, sexe, âge, minorités). Veuillez vous référer au paragraphe « Risques, y compris l'éthique et la confidentialité » de la section « Exposé de la proposition » du présent document et examiner attentivement le sujet de l'éthique relative à la collecte de données.

Vous pouvez consulter les ensembles de données provenant de projets sélectionnés dans le cadre des appels à propositions sur le TALN du Lacuna Fund pour 2020 et 2021 afin de prendre connaissance des travaux en cours.

# 4 – Informations sur les propositions

**Remarque : Le site internet du Lacuna Fund comprend de nombreuses ressources, telles que des références pertinentes sur la qualité des données, la documentation et le format, pour aider les candidats à soumettre une proposition compétitive.**

**Les propositions seront acceptées uniquement via le portail SurveyMonkey Apply disponible à l'adresse www.lacunafund.org/fr/soumettre-un-projet/**. Les candidatures peuvent être soumises en anglais, espagnol, français et portugais. Une description des questions relatives aux candidatures est proposée ci-après à titre d'information uniquement. Les sections suivantes doivent obligatoirement être complétées :

- Informations sur le candidat (accessibles dans le portail)
- Exposé de la proposition
- Budget et calendrier

## Informations sur le candidat

Cette section invite le candidat à fournir :

- Un résumé de la proposition en 200-250 mots ;
- Des détails sur l'institution (ou les institutions) et/ou l'équipe candidate ;

- Le lieu des travaux ;
- Le CV des principaux membres de l'équipe ;
- Les processus d'examen éthique de l'institution ou des institutions affiliées ;
- La capacité de l'équipe à obtenir des approbations nationales.

## Exposé de la proposition

*Veuillez limiter votre exposé de proposition à 10 pages, références non incluses, avec des marges de 2,5 cm et une police de caractères de 11 points minimum. Les annexes ou les exposés de proposition de plus de 10 pages ne seront pas examinés.*

Dans cette section, le candidat est invité à télécharger un exposé cohérent, au format PDF ou Word, qui traite des points suivants :

- **Qualifications** – Veuillez décrire la ou les organisation(s) ou le ou les partenariat(s) présentant la proposition, comment ils satisfont aux critères d'éligibilité énoncés ci-dessus, et vos qualifications uniques pour entreprendre le travail proposé.
- **Identification du problème, solution proposée et ensemble de données** – Veuillez décrire le problème ou la lacune dans les données de formation ou d'évaluation et la solution proposée. Veuillez résumer le ou les ensembles de données que vous avez l'intention de créer, développer, consolider ou tenir à jour. *Veuillez expliquer comment votre projet comble une lacune et complète les travaux existants.*
- **Spécifications et livrables pour les données et la documentation proposées** – Veuillez inclure les éléments suivants :
  - La quantité de données qui sera incluse dans l'ensemble de données.
  - Les types et le format des données et/ou des annotations, ainsi que le cadre et la taille de l'échantillon ou un plan pour assurer la représentation, le cas échéant.
  - Les mesures à utiliser pour évaluer les résultats souhaités de la création de données (par exemple, les mesures d'équité dans l'ensemble de données, un QA/QC par rapport à un point de référence, etc.)
- **Bénéficiaires visés et cas d'utilisation** – Veuillez décrire les consultations antérieures et/ou la collaboration proposée avec les bénéficiaires visés et décrire les potentiels cas d'utilisation actuels et futurs de l'apprentissage automatique pour les ensembles de données proposés. Veuillez expliquer comment l'ensemble de données respecte et reflète la diversité des communautés qu'il représente. Veuillez indiquer comment la qualité proposée, les méthodes de collecte et d'autres détails font que les données peuvent être utilisées dans ce contexte opérationnel particulier.
- **Méthodologie** – Veuillez fournir un bref aperçu des étapes proposées (et des hypothèses clés) pour élaborer et mettre en œuvre le projet. Veuillez inclure :
  - Les techniques proposées de collecte et d'annotation des données et les informations sur l'interopérabilité. Veuillez tenir compte de l'infrastructure existante ou commune et des dernières techniques et outils pour accélérer la collecte, le nettoyage et le partage des données.

- Les mesures de contrôle de la qualité, telles que la qualité que tous les échantillons de données doivent atteindre pour l'ensemble de données. Veuillez indiquer comment l'équipe prévoit de traiter les valeurs aberrantes qui pourraient affecter la qualité de l'ensemble de données.
- Un plan pour évaluer et atténuer les erreurs et les biais (par exemple, les préjugés sexistes ou autres).
- Les modalités prévues d'exploitation des ressources existantes, y compris les méthodes ou technologies de collecte, en reliant les ensembles de données préexistants sur le domaine, ainsi que les ressources existantes dans d'autres contextes.
- Les autorisations obtenues ou les mesures que vous prendrez pour obtenir les autorisations nationales ou autres autorisations requises. Déterminez quelles juridictions exigent des approbations et si la recherche proposée correspond à la définition de la recherche dans cette juridiction. Si vous estimez que les approbations locales, nationales ou régionales ne sont *pas* nécessaires, veuillez expliquer pourquoi.
- Tout défi ou incertitude anticipé dans la collecte des données et les contre-mesures proposées.

- **Impact transformationnel** – Veuillez expliquer comment l'ensemble de données proposé, ou l'annotation, contribuera à produire l'impact souhaité. Le cas échéant, décrivez comment les produits peuvent motiver une utilisation dans les contextes de recherche ou commerciaux. Notez toute contrainte pratique à laquelle vous pourriez être confronté (par exemple, la pénétration d'Internet).
- **Gestion des données et licences** – Veuillez décrire :
  - Tout problème anticipé lié au droit d'auteur pour les données sources et la collaboration avec le détenteur des droits. Veuillez également aborder tous les problèmes prévus en matière de droits d'auteur et d'octroi de licences pour les données secondaires.
  - Les plans pour des licences favorisant une utilisation responsable en aval. Conformément à la politique de propriété intellectuelle du Lacuna Fund, l'ensemble de données et toute propriété intellectuelle connexe, telle que les méthodes de collecte, les feuilles de données, la façon de charger ou de lire les ensembles de données, ou toute autre information permettant d'assurer la facilité d'utilisation, doivent être mis à disposition sous une licence de source ouverte, par attribution (CC-BY 4.0 ou CC BY-SA 4.0). Si une licence plus restrictive est proposée, il convient d'en donner la raison. Le budget peut inclure des ressources pour les licences.
  - Si vous avez l'intention d'utiliser un ensemble de données existant pour votre projet, veuillez indiquer que votre équipe a reçu les autorisations nécessaires de la part du propriétaire de l'ensemble de données afin que celui-ci puisse être diffusé conformément à la politique de propriété intellectuelle du Lacuna Fund, ou justifiez une autre structure de licence. *Les lettres de soutien de titulaires de données existants sont facultatives mais encouragées.*
  - Prévoyez d'inclure un fichier de métadonnées et une feuille de données comme documentation pour votre ensemble de données, conformément aux orientations décrites dans le document du Lacuna Fund intitulé Dataset Hosting and Documentation Guidance (Orientations en matière d'hébergement des données et de documentation).

- La plateforme d'hébergement que vous comptez utiliser. Les plateformes d'hébergement doivent attribuer un identifiant numérique d'objet (DOI) à l'ensemble de données, quantifier les téléchargements de l'ensemble de données et collecter les informations de contact pour les téléchargements de l'ensemble de données. Pour obtenir des orientations sur les plateformes d'hébergement suggérées, voir le document Dataset Hosting and Documentation Guidance (Orientations en matière d'hébergement des données et de documentation).

- **Risques, y compris l'éthique et la confidentialité** – Veuillez identifier les risques potentiels, y compris, mais sans s'y limiter, les problèmes potentiels de confidentialité et d'éthique, et décrivez les mesures que vous prendrez pour les atténuer. Plus précisément :
  - Une réflexion sur les communautés dont vous êtes issu et sur vos identités, et sur l'impact qu'elles peuvent avoir sur votre travail.
  - Indiquez comment vous garantirez le consentement éclairé, le cas échéant (ceci doit inclure la notification des cas potentiels d'utilisation future des données).
  - Décrivez comment vous garantirez l'équité tout au long du projet, y compris, mais sans s'y limiter, une rémunération équitable pour l'étiquetage, l'annotation, la fourniture de données, et d'autres services d'IA. Voici quelques pistes pour donner la priorité à des normes de travail et à des pratiques de rémunération équitables :
  - Suivre et signer la charte des normes du travail équitable (Fair Work) pour les travaux d'IA lorsque l'annotation et la fourniture de données sont sous-traitées à des entités commerciales.
  - Suivre une approche de fourniture et d'annotation des données orientée vers la communauté, si un tel modèle existe.
  - Présenter une dérogation volontaire/une enquête aux participants au projet afin de remédier à toute charge indue.
  - Décrivez comment la diversité de genre et d'autres considérations démographiques sont intégrées dans l'équipe de projet, la collecte de données de formation et le développement de modèles, afin que les ensembles de données représentent fidèlement les impacts sur les différentes communautés / différents groupes.
  - Présentez un plan pour l'anonymisation des informations personnellement identifiables et la conformité avec les lois sur la confidentialité, le cas échéant. Si aucun cadre juridique national n'est disponible, la proposition doit décrire ou faire référence aux bonnes pratiques. Veuillez incorporer les considérations en matière de respect de la confidentialité des informations au niveau des individus et des communautés. Veuillez vous référer au document Data Anonymization Guide (Guide de l'anonymisation des données) du Lacuna Fund.
  - Discutez des impacts négatifs potentiels dans la production et l'utilisation de l'ensemble de données et des mesures pour les atténuer, y compris des risques potentiels pour les droits de l'homme et la possibilité d'une forte consommation d'énergie par la technologie de l'IA conduisant à une augmentation des émissions de dioxyde de carbone.

- **Planification de la durabilité et de la communication**
  - Veuillez décrire comment l'ensemble de données annoté sera maintenu et/ou étendu au-delà du financement initial (par exemple, par le biais d'un modèle de référence,

d'applications d'AM résultantes, par une communauté dédiée, ou un groupe de parties intéressées avec un modèle de gouvernance robuste pour l'ensemble de données ouvert) et comment un cas d'utilisation potentiel pourrait soutenir le projet.

- Veuillez expliquer comment l'ensemble de données respectera les principes des données FAIR (https://www.go-fair.org/resources/faq/what-is-fair/). Indiquez les mesures que vous prendrez pour vous assurer que l'ensemble de données est **facile à trouver, accessible, interopérable et réutilisable**.
- Veuillez décrire les activités de communication visant à faire connaître le ou les ensembles 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 pérennité de l'ensemble de données avec les parties prenantes intéressées ou de la mise en place d'un comité de pérennité.

## Calendrier du projet et livrables

Dans cette section, le demandeur est invité à soumettre un tableau indiquant le calendrier d'achèvement des principales activités et des livrables. Le calendrier peut inclure, sans s'y limiter, la formation du personnel, la collecte des données, l'annotation, l'assurance qualité, la validation/le nettoyage et la publication des données. Les livrables peuvent inclure, sans s'y limiter, des parties de l'ensemble de données pour montrer l'avancée du projet, l'ensemble de données complet et la documentation d'accompagnement ou les méthodes de collecte à mettre en libre accès.

*Tous les délais doivent inclure une date à laquelle les données seront accessibles au public avec toute la documentation.*

**Remarque** : les projets proposés doivent être achevés, les ensembles de données publiés et les rapports finaux soumis au plus tard en 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.

## Budget

Veuillez fournir 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 Lacuna Fund, disponible sur le portail des candidats.

La dotation totale disponible est d'environ 1 million de dollars (USD). Nous aimerions financer des projets dans chacune des régions ciblées (Afrique, Amérique latine) et nous pensons pouvoir soutenir 6-8 petits projets avec des budgets jusqu'à 100 000 dollars et 2 à 3 projets plus gros et plus complexes avec des budgets de l'ordre de 100 000 à 250 000 dollars. Le Groupe 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. Les budgets peuvent inclure, mais ne sont pas limités à, des coûts pour :

- le renforcement des capacités liées à la collecte de données et à l'assurance qualité/contrôle de la qualité,

- la collecte des données (y compris une rémunération équitable pour la fourniture de données),
- l'annotation des données (y compris une rémunération équitable pour l'étiquetage de données),
- l'AQ/CQ ou la vérification,
- le post-traitement des données,
- la publication des données,
- l'octroi de licences,
- la publication des résultats en libre accès,
- le temps nécessaire pour préparer une déclaration de données pour l'ensemble de données,
- les efforts de crowdsourcing, tels que les « label-a-thons »,
- le stockage de données,
- la puissance de calcul,
- les ateliers,
- les activités de communication, y compris la participation à deux conférences au maximum pour présenter les ensembles de données.

Les fonds ne peuvent pas être utilisés pour le paiement direct de droits de douane, d'importation ou d'autres droits ou taxes prélevés à l'importation de biens ou d'équipements dans un pays ou une juridiction quelconque. **Les taux indirects sont strictement limités à 14.5 % des coûts de recherche directs.**

Les partenaires du Lacuna Fund peuvent fournir des services de stockage dans le cloud et de la puissance de calcul en nature. Si vous souhaitez utiliser cette ressource, veuillez l'inclure dans votre budget. Les équipes sélectionnées recevront des instructions sur la manière de postuler pour ces services lorsqu'elles recevront leur subvention.

Voir la feuille d'instructions dans le modèle de budget pour plus d'informations sur les directives budgétaires, y compris les informations sur les frais de personnel admissibles.

Nous vous remercions de l'intérêt que vous portez au Lacuna Fund et de vos efforts visant à soutenir une plus grande équité et accessibilité des applications d'AM à l'appui du traitement automatique du langage naturel (TALN). Nous avons hâte d'examiner vos propositions !

# REMARQUE : possibilités de mentorat et de mise en relation

Lacuna Fund est heureux de s'associer à la Masakhane Research Foundation (MRF) pour offrir des possibilités de mentorat et de mise en relation aux candidats dans le domaine du TALN. La MRF est une organisation de terrain dont la mission est de renforcer et de stimuler la recherche en matière de TALN dans les langues africaines, pour les Africains, par les Africains. L'objectif de la MRF est de permettre aux Africains de concevoir et de s'approprier ces avancées technologiques visant à favoriser la dignité humaine, le bien-être et l'équité, grâce à la création de communautés inclusives, à la recherche participative ouverte et à la pluridisciplinarité.

Pour cet appel à propositions du Lacuna Fund, la MRF offrira aux candidats d'Afrique et d'Amérique latine une occasion unique de se joindre à la communauté de la MRF et mettra en relation les candidats intéressés avec un mentor qui examinera le projet de proposition et discutera des moyens de l'améliorer. Les personnes intéressées peuvent demander une session avec un mentor en remplissant ce formulaire Google avec 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 du TALN dans les langues à faibles ressources », « rédiger une proposition de recherche »*** et ***« préparer des budgets »***. En fournissant une description détaillée, vous aiderez les candidats à trouver des mentors qui correspondent à votre domaine d'intérêt. La mise en correspondance des mentors et des mentorés sera gérée par nos partenaires associés en la matière, qui organiseront plusieurs ateliers pour réunir les candidats qui travaillent dans des domaines connexes (par exemple, les linguistes et les chercheurs en TALN) et coordonneront et organiseront des sessions de communication entre les groupes concernés. Dans le cadre de ce processus, nous offrirons également aux candidats la possibilité de collaborer entre eux afin de former des équipes de projet et de soumettre des propositions ensemble.

Les candidats intéressés sont encouragés à postuler pour une session de mentorat au moins 6 semaines avant la date d'échéance de la proposition, au plus tard le **15 juillet 2024**. Nous ferons de notre mieux pour identifier un mentor pour chaque personne qui en fait la demande, mais nous ne pouvons pas garantir que des mentors seront disponibles pour tous. 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.