Appel à manifestation d'intérêt : Ensemble de données d'apprentissage machine pour les langues en Afrique subsaharienne - 2021

Lacuna Fund: Our Voice on Data

19 Octobre 2021

Table des matières

1 – Introduction

Objet et objectifs du Fonds

Le Lacuna Fund soutient la création, le développement et la tenue à jour d'ensembles de données d'évaluation et de formations équitables permettant une application fiable des outils d'apprentissage machine présentant une valeur sociale élevée.

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 l'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 le développement et la tenue à jour d'ensembles de données étiquetées dans le respect des principes d'équité.

Principes du Fonds

Les principes suivants guideront la gouvernance et les opérations du Lacuna Fund.

- **Accessibilité** – le Lacuna Fund s'engage à garantir que les ensembles de données créés grâce à son financement sont accessibles aux communautés mal desservies et leur bénéficient au service des objectifs énoncés précédemment. Les ensembles de données et la propriété intellectuelle dont ils font l'objet seront couverts par des licences de données ouvertes appropriées pour favoriser une utilisation responsable en aval. (Voir la Politique en matière de propriété intellectuelle pour plus de détails.)
- **Équité** – le Lacuna Fund vise à rendre l'IA plus équitable en favorisant des ensembles de données qui sont créés par des personnes aux profils sous-représentés dans le monde et qui répondent à leurs besoins. Ces ensembles de données ne doivent pas créer ou renforcer des biais ou préjugés (par exemple, ils doivent être inclusifs en termes de genre et représentatifs des personnes de couleur à l'échelle mondiale) ni soutenir des systèmes ou des technologies susceptibles de nuire.
- **Éthique** – le Lacuna Fund financera la collecte de données d'une manière conforme aux normes éthiques du travail et exigera des bénéficiaires de subventions qu'ils précisent les mesures qu'ils prendront pour protéger la vie privée et éviter tout préjudice dans la collecte, l'octroi de licences et l'utilisation des ensembles de données créés avec les fonds des subventions.
- **Approche participative** – le Lacuna Fund s'efforce de répondre aux besoins des parties prenantes concernées en encourageant le leadership ou un engagement fort des experts locaux, des bénéficiaires et des utilisateurs finaux dans la gouvernance du Fonds et dans les

- **Qualité** – les données produites grâce aux efforts financés par Lacuna doivent être d'une qualité élevée permettant des applications bénéfiques pour la recherche, les communautés et l'industrie.
- **Impact transformationnel** – en finançant des ensembles de données qui comblent des lacunes fondamentales de l'IA, le Lacuna Fund entend libérer les avancées dont celle-ci est porteuse au bénéfice des communautés pauvres et mal desservies.

Philosophie en matière d'octroi de subventions

Le 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.

Le 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 (GCT), le 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.**

Le Lacuna Fund est soutenu par la Fondation Rockefeller, Google.org et le Centre de recherche pour le développement international du Canada. Par ailleurs, cet appel relatif à des ensembles de données étiquetées dans le respect des principes d'équité pour de meilleurs résultats en matière de santé est soutenu par les fondations Wellcome Trust et The Gordon and Betty Moore Foundation.

2- Vue d'ensemble

Critères d'éligibilité applicables aux organisations

Le 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, mais seul le candidat principal recevra des fonds ;
- avoir une mission de soutien du bien sociétal, au sens large ;
- être une organisation dont le siège se situe dans le pays ou la région dans lequel les données seront collectées ou avoir un partenariat important avec celui-ci/celle-ci ;

- disposer de toutes les autorisations requises, nationales ou autres, pour poursuivre les recherches proposées, ainsi que des accords d'utilisation des données ou des projets pour les obtenir. 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 de la faculté de développer cette capacité grâce au partenariat décrit dans la manifestation d'intérêt – pour procéder à l'étiquetage, à la création, à la consolidation, 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. les résultats en matière de soins de santé) afin de permettre à plusieurs entités de réaliser des travaux analytiques de haute qualité pour l'IA/l'apprentissage machine.

Procédure de sélection et critères d'évaluation

Le Lacuna Fund est à la recherche de manifestations d'intérêt d'organisations désireuses de débloquer, créer, consolider et/ou améliorer des ensembles de données de formation et d'évaluation prenant en charge le traitement automatique du langage naturel pour les langues d'Afrique subsaharienne. Le Lacuna Fund et ses partenaires procéderont à une sélection initiale des manifestations d'intérêt pour vérifier l'éligibilité organisationnelle et la faisabilité du projet. Les membres du Groupe consultatif technique ne peuvent pas soumettre une manifestation d'intérêt ou 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 à propositions évaluera les manifestations d'intérêt pour sélectionner les quelques organisations qui seront invitées à présenter des propositions complètes pour financement. Les candidats invités seront sélectionnés sur la base du niveau de satisfaction des critères suivants :

- **Qualité** – l'organisation ou l'équipe proposant le projet démontre une collaboration interdisciplinaire entre des experts qualifiés dans les langues, la recherche, l'apprentissage machine et la gestion de données. Les partenariats entre les acteurs bien dotés en ressources ayant de bonnes capacités de recherche et des acteurs servant des populations marginalisées et des langues à faibles ressources sont encouragés.
- **Impact transformationnel** – les ensembles de données doivent améliorer l'apprentissage machine et impliquer des cas d'utilisation qui procurent un avantage social tangible aux populations mal desservies. Les exemples incluent, sans s'y limiter : a) l'étiquetage, le nettoyage ou la collecte de données de validation ou la mise en commun d'ensembles de données existants pour dégager une valeur supplémentaire ou assurer une plus grande précision de l'ensemble de données existant ; b) la création d'un nouvel ensemble de données étiquetées de grande valeur pour une langue faiblement dotée en ressources ; c) l'amélioration de la représentativité d'un ensemble de données existant et de la prise en compte de la race, du genre, de l'ethnie, des capacités, etc.
- **Équité** – la proposition est assortie d'un argumentaire convaincant démontrant que les ensembles de données seront mis en œuvre afin d'aider les populations vulnérables et mal desservies.

- **Approche participative** – les ensembles de données doivent être centrés sur les besoins des communautés concernées et il convient de travailler avec les partenaires pour déterminer les avantages pour la communauté. Plus précisément, les projets doivent associer les membres de la communauté aux décisions relatives à la gouvernance des données (par exemple, quelles données sont conservées et comment elles sont utilisées). Lorsque l'ensemble de données a une portée géographique, l'équipe est majoritairement implantée dans la région concernée et/ou entretient des liens étroits avec les acteurs locaux pour garantir une tenue à jour et un usage pérennes de l'ensemble de données par la communauté locale.
- **Éthique** – le projet 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.
- **Efficacité** – le porteur de projet a tenu compte des ensembles de données existants 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.
- **Faisabilité** – le projet est réalisable en fonction du budget et de l'étendue des travaux proposés. Bien que la manifestation d'intérêt ne doive pas inclure un budget détaillé, nous nous attendons à ce que les manifestations d'intérêt incluent une estimation générale du budget global prévu pour mener à bien le projet.
- **Accessibilité** – les ensembles de données seront : a) largement accessibles dans le cadre d'une licence ouverte conformément à la Politique du Lacuna Fund en matière de propriété intellectuelle. 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, ainsi qu'un mécanisme pour fournir l'accès en vertu de la licence proposée.
- **Durabilité** – le projet comporte un plan visant à garantir la durabilité et la tenue à jour future de l'ensemble de données, par exemple par une communauté spécifique ou un groupe de parties intéressées (organisations à but lucratif ou non lucratif), et un modèle de gouvernance robuste pour l'ensemble de données ouvert.

Calendrier

Appel à manifestation d'intérêt diffusé publiquement sur le site web du Lacuna Fund	**19 octobre 2021**
Date limite pour les questions/réponses Veuillez adresser vos questions à secretariat@lacunafund.org	**2 novembre 2021**
Publication des réponses	**9 novembre 2021**
Date limite pour les manifestations d'intérêt	**1er décembre 2021**

Période de questions/réponses : toutes les questions concernant la manifestation d'intérêt doivent être adressées par courrier électronique à secretariat@lacunafund.org en précisant « Question Manifestation d'intérêt Langues 2021 » dans l'objet. Les questions adressées d'ici le 5 novembre 2021 seront anonymisées et leur réponse sera publiée le 15 novembre 2021 dans un document posté sur la page « Soumettre un projet » du site web de Lacuna Fund.

3 – Objet et besoins

Objet

Le but de cet appel à manifestation d'intérêt est de répertorier les projets qui seront invités à soumettre des propositions complètes pour l'élaboration d'ensembles de données ouverts et accessibles pour les applications d'apprentissage machine qui permettront le traitement du langage naturel pour les langues d'Afrique subsaharienne. La capacité à communiquer et à être compris dans sa propre langue est indispensable à l'inclusion numérique et sociétale. Les techniques de traitement automatique du langage naturel (TALN) ont permis de développer des applications décisives de l'IA qui facilitent l'inclusion numérique et les améliorations dans de nombreux domaines, notamment l'éducation, la finance, les soins de santé, l'agriculture, la communication et la réponse aux catastrophes, entre autres. De nombreuses avancées dans le domaine du TALN, tant fondamental qu'appliqué, ont été réalisées à partir d'ensembles de données sous licence ouverte et accessibles au public.

Cependant, ces ensembles de données ouverts et accessibles au public sont rares, voire inexistants, pour de nombreuses langues africaines, ce qui signifie que les avantages du TALN ne sont pas accessibles aux locuteurs de ces langues. Quand des ensembles de données pertinents existent, ils s'appuient souvent sur des textes religieux, missionnaires ou judiciaires, produisant alors une langue désuète et créant des préjugés. Il est nécessaire de disposer d'ensembles de données issus de texte, de discours et autres, librement accessibles, afin de faciliter les progrès fondés sur les technologies du TALN pour les langues africaines.

Besoins

Le Lacuna Fund lance un appel à manifestation d'intérêt de la part d'organisations qualifiées désireuses d'élaborer des ensembles de données de formation et d'évaluation ouverts et accessibles pour des applications d'apprentissage machine dans le domaine du TALN en Afrique subsaharienne. Le GCT reconnaît l'importance des ensembles de données qui auraient un impact significatif quel que soit le nombre de locuteurs de la langue concernée, ainsi que le besoin d'ensembles de données multilingues.

Les appels à manifestation d'intérêt peuvent inclure, sans s'y limiter :

- la collecte et/ou l'annotation de nouvelles données ;
- l'annotation ou la publication de données existantes ;
- une augmentation des ensembles de données existants dans tous les domaines pour réduire les préjugés (comme les préjugés sexistes ou d'autres types de préjugés ou de discriminations) ou une amélioration de la facilité d'utilisation des technologies de TALN dans des contextes de revenus faibles ou moyens ;

- la création de petites données de référence de qualité supérieure pour les tâches de TALN dans les langues africaines à faibles ressources.

Bien que le Lacuna Fund se concentre principalement sur la création, l'annotation, l'augmentation et la tenue à jour des ensembles de données, les **propositions peuvent inclure le développement d'un modèle de référence** pour garantir la qualité de l'ensemble de données financé et/ou pour faciliter l'utilisation de l'ensemble de données pour des applications bénéfiques sur le plan social.

Le GCT reconnaît le besoin d'avoir des ensembles de données qui permettent une meilleure exécution des tâches clés du TALN dans les langues africaines, ainsi que l'évaluation de la performance des systèmes dans les langues africaines, **notamment mais pas uniquement, les éléments suivants** :

- les **corpus oraux**, notamment pour les applications qui permettent aux analphabètes ou autres groupes défavorisés d'accéder aux outils, informations et/ou services technologiques. Plus particulièrement, il est nécessaire de constituer de vastes corpus oraux de haute qualité (c'est-à-dire représentatifs de la population, phonétiquement riches et équilibrés, ayant une orthographe propre, transcriptions orthographiques précises, sans paraphrase ni suppression des disfluences) ou des ensembles de données destinés à soutenir les systèmes de TALN sans texte ;
- les **corpus textuels étiquetés** à utiliser comme données de formation ou d'évaluation de référence, y compris les corpus parallèles pour la traduction automatique ou les corpus destinés à d'autres tâches fondamentales ou en aval du TALN. Les tâches en aval peuvent inclure, sans toutefois s'y limiter : la réponse aux questions et l'IA conversationnelle, les ensembles de données d'analyse de sentiments ; le résumé automatique de texte ou d'autres tâches de compréhension et de génération de langage naturel, ou des ressources pour soutenir la formation au TALN ;
- les **corpus textuels non étiquetés** pour les modèles de langue qui offrent de multiples pistes de recherche ou d'application. Cela inclut les corpus textuels qui peuvent être utilisés à l'appui de la formation et de l'évaluation des modèles de parole ;
- les **ensembles de données relatifs à des textes ou des discours avec alternance codique** qui améliorent les performances des tâches de TALN dans de telles situations ;
- la **création ou l'augmentation d'ensembles de données de textes ou de discours propres à un domaine**, comme des ensembles de données numériques, des noms d'endroits ou des combinaisons de mots ou des phrases spécifiques qui permettent des applications avec un impact social important ;
- les **autres ensembles de données multimodaux et innovants**, comme le sous-titrage audio ou vidéo ou d'autres interactions image-texte.

Pour tous les projets proposés, le GCT encourage les équipes à :

- exploiter les ressources existantes, notamment les méthodes ou outils de collecte ainsi que les corpus ou modèles existants dans les langues à ressources plus élevées ;
- favoriser une collaboration et une contribution solides des linguistes et des chercheurs en éthique de l'IA au sein des équipes de projet ;

- considérer les éléments sociologiques du TALN (par exemple, les différences dans la description verbale d'une même image selon les cultures) ;
- s'étendre au-delà des langues les plus courantes et les plus populaires pour développer la communauté ;
- inclure les variations régionales et autres dans les ensembles de données.

Découvrez les projets sélectionnés dans le cadre de notre cycle de financement 2020 ici pour voir quels travaux sont actuellement en cours. Nous encourageons les projets qui adoptent des approches totalement nouvelles et abordent des langues différentes ou qui s'inspirent de ces projets soutenus.

4 – Informations sur la manifestation d'intérêt

La manifestation d'intérêt doit être soumise exclusivement via le portail spécifique disponible à l'adresse www.lacunafund.org/fr/soumettre-un-projet/. Une description des questions relatives aux candidatures est proposée ci-après à titre d'information uniquement. ***Veuillez limiter votre manifestation d'intérêt à 4 pages, références non incluses****, avec des marges de 2,5 cm et une police de caractères de 11 points minimum. Les annexes ou supports de la manifestation d'intérêt au-delà des 4 pages ne seront pas examinés.*

Les informations suivantes doivent y figurer :

- **Descriptions et qualifications des équipes de projet participantes** – veuillez fournir un bref aperçu de la mission des organisations participantes, des services fournis et des groupes d'intérêt desservis, de la manière dont elles satisfont aux critères d'éligibilité énoncés ci-dessus, et des qualifications particulières du candidat pour entreprendre le travail proposé, y compris son expérience dans l'élaboration et le partage d'ensembles de données sur les langues.
- **Description et spécifications du projet** – veuillez résumer brièvement l'ensemble de données que vous souhaitez créer, augmenter ou tenir à jour et le besoin spécifique en matière d'apprentissage machine que l'ensemble de données satisferait dans le contexte de la ou des langues particulières ou également des tâches de TALN. Si possible, veuillez fournir des informations spécifiques sur l'échelle (par exemple, heures ou phrases de données) et le contenu (par exemple, format et contenu des étiquettes, le cas échéant) de l'ensemble de données proposé.
- **Cas d'utilisation et avantages prévus** – quels cas d'utilisation pertinents l'ensemble de données permettra-t-il ? Veuillez décrire le contexte social et culturel des technologies linguistiques qui pourraient en résulter et, le cas échéant, décrire comment l'ensemble de données, les modèles et les produits connexes pourraient encourager et permettre des voies d'application multiples et durables. Le Groupe consultatif technique s'intéresse aux ensembles de données qui font progresser l'état de la recherche et de la pratique du traitement automatique du langage naturel africain, ainsi qu'à ceux qui permettent des applications bénéfiques sur le plan social.
- **Méthodologie du projet** – veuillez expliquer comment les données seront collectées, étiquetées ou augmentées. Veuillez décrire les mesures que le projet prendra pour assurer la représentation et éviter les biais, ainsi que pour garantir la qualité des données collectées

ou étiquetées. Si les données sont destinées à une externalisation à grande échelle, veuillez fournir des informations sur les communautés engagées existantes, les incitations ou la faisabilité de l'approche.

- **DPI, éthique et confidentialité** – veuillez confirmer que l'ensemble de données proposé et la propriété intellectuelle connexe feront l'objet d'une licence internationale CC-BY 4.0 conformément à la politique de propriété intellectuelle du Lacuna Fund. Si un système d'octroi de licence plus restrictif est proposé, veuillez fournir une justification. Si le projet prévoit d'utiliser ou de consolider des ensembles de données existants, des corpus ou d'autres ressources, veuillez fournir des informations sur la disponibilité et les licences.
- **Calendrier général et budget global pour la mise en œuvre du projet** :
  - Calendrier – veuillez proposer un calendrier général pour l'achèvement des étapes incluses dans la méthodologie générale ci-dessus, y compris le nombre total de mois nécessaires à l'achèvement.
  - Budget – veuillez fournir une vue d'ensemble du budget prévu pour la réalisation des étapes incluses dans la méthodologie générale ci-dessus. Les budgets doivent être présentés en dollars américains (USD). La dotation totale disponible est d'environ 900 000 USD. Nous prévoyons des budgets dans une fourchette de 10 000 à 100 000 USD pour les projets de petite taille à taille moyenne, et des budgets pouvant atteindre 200 000 USD pour les projets de grande ampleur et complexes. Nous pensons pouvoir financer deux ou trois grands projets et un grand nombre de petits projets. **Les coûts indirects sont limités à 12 % du budget proposé.**

Nous vous remercions de l'intérêt que vous portez au Lacuna Fund et de vos efforts visant à développer des applications d'apprentissage machine dans le domaine du traitement automatique du langage naturel. Nous avons hâte d'examiner vos propositions !