Conseils pour l'hébergement et la documentation des ensembles de données

Lacuna Fund : Notre voix sur les données

Février 2022 ; Mis à jour en février 2023

Ce document fournit des conseils aux bénéficiaires de subventions concernant l'hébergement, la documentation et l'octroi de licences pour les ensembles de données créés avec le soutien du Lacuna Fund. Il s'agit d'un document évolutif ; veuillez faire part de tout ajout ou suggestion d'amélioration dans les commentaires du document.

Accessibilité et hébergement

La politique d'hébergement de Lacuna Fund est flexible afin de permettre la propriété locale des données et de répondre aux cas d'utilisation envisagés pour les données. Conformément aux principes de Lacuna Fund, l'hébergement des jeux de données doit être largement accessible et permettre une utilisation socialement bénéfique des données.

Lacuna Fund exige des bénéficiaires de subventions qu'ils hébergent leurs données sur un site répondant aux conditions suivantes, afin que les ensembles de données puissent être trouvés et que leur utilisation puisse être suivie :

- Attribue un identifiant d'objet numérique (DOI) aux ensembles de données ou permet d'en joindre un dans le cadre des métadonnées.
- Est indexé par les principaux moteurs de recherche (par exemple, Google Dataset Search ou des outils similaires).
- Est fiable et persévérant.
- Quantifie le nombre de consultations et de téléchargements de la page de destination pour l'ensemble de données.
- Collecte des informations de contact pour les téléchargements d'ensembles de données d'une manière qui maximise la conversion.

Envisagez des solutions d'hébergement qui permettent l'intégration d'outils couramment utilisés pour traiter les données, qui sont déjà utilisées par les communautés susceptibles de mettre en œuvre les cas

d'utilisation envisagés, et qui permettent une planification solide de la durabilité ou un modèle de gouvernance pour les données.

Des informations supplémentaires et des exemples de plateformes d'hébergement potentielles sont disponibles à l'annexe A.

Documentation du jeu de données

Lacuna Fund reconnaît qu'une documentation claire est essentielle pour garantir l'accessibilité et l'utilisation du jeu de données. **Lacuna Fund demande aux bénéficiaires de subventions d'inclure la documentation suivante lors de la soumission des jeux de données :**

1) Fichier de métadonnées (voir l'annexe B pour un modèle)
2) Fiche technique (voir annexe C pour les modèles)
3) Identifiant d'objet numérique (DOI)

Les liens suivants fournissent des normes et des ressources générales et spécifiques à un domaine :

- Normes pour les métadonnées - par exemple, le catalogue des biens spatio-temporels (STAC) pour les données liées à l'observation de la terre.
- Documentation normalisée pour les ensembles de données de ML, comme les feuilles de données pour les ensembles de données (Gebru et. al.)
  - Pour les modèles types, Model Cards for Model Reporting (Mitchell et. al)
  - Pour le NLP, voir Data Statements for Natural Language Processing (Bender et Friedman).
- Taxonomies et autres ontologies à intégrer qui favorisent l'interopérabilité et l'utilisabilité des ensembles de données.

Licences

L'octroi de licences pour les ensembles de données doit se faire conformément à la politique de propriété intellectuelle du Lacuna Fund, qui stipule que : "Les ensembles de données et la propriété intellectuelle connexe développés avec des fonds de subvention seront : (a) appartiendront à l'entité bénéficiaire de la subvention ; et (b) feront l'objet d'une licence ouverte par l'entité bénéficiaire de la subvention afin de maximiser le potentiel de diffusion responsable et d'utilisation en aval de cette propriété intellectuelle. Le bénéficiaire de la subvention donnera la priorité à la diffusion de la propriété intellectuelle dans le cadre d'une structure de licence de source ouverte permissive telle que Apache 2.0 (https://opensource.org/licenses/Apache-2.0) pour tout code ou autres inventions, ou CC-BY 4.0 International (https://creativecommons.org/licenses/by/4.0/) pour toute autre propriété intellectuelle (par exemple, les œuvres créatives qui ne sont pas du code ou brevetables)."

" La propriété intellectuelle connexe " comprend généralement, mais sans s'y limiter, les éléments suivants :

- Métadonnées, fiches de données ou autres informations.
- Des conseils sont inclus avec l'ensemble de données pour guider son utilisation.

- Exemples de modèles ou autres outils d'information publiés avec les données.

Les groupes consultatifs techniques peuvent exercer un pouvoir discrétionnaire raisonnable concernant des structures de licence plus restrictives afin de protéger la vie privée, d'éviter tout préjudice ou de maximiser le potentiel de diffusion responsable et d'utilisation en aval. Les **demandes d'exception à la licence internationale CC-BY 4.0 pour les ensembles de données doivent être faites au stade de la proposition**. Si une exception est accordée, elle sera intégrée à votre contrat. Si les circonstances de votre projet ont subi des changements importants qui nécessiteraient une licence différente, veuillez en informer le secrétariat dès que possible.

Annexe A : Options d'hébergement et conseils supplémentaires

Veuillez communiquer avec l'organisation qui hébergera votre ensemble de données au début de votre projet afin de comprendre les délais de publication de l'ensemble de données et les coûts potentiels à partager, le cas échéant, en particulier pour les sites d'hébergement qui conservent et téléchargent manuellement les données. Des informations sur le processus que vous entreprendrez pour publier vos données seront demandées dans le rapport à mi-parcours.

Les subventions du Lacuna Fund exigent que votre ensemble de données soit disponible publiquement et ouvertement à la fin de la période d'exécution. Si la présentation à une conférence ou la publication d'un travail universitaire risque d'entraîner un retard dans la mise à disposition de votre ensemble de données, envisagez un serveur de préimpression ou un site dédié avant l'hébergement définitif de l'ensemble de données, comprenant des informations claires sur les licences et la gouvernance des données.

La publication dans une revue universitaire n'est PAS obligatoire. Toutefois, si vous publiez un article lié à votre ensemble de données, nous serions ravis d'en entendre parler !

Les options d'hébergement suivantes répondent à tout ou partie des exigences décrites dans les conseils principaux ci-dessus. Il est possible d'héberger vos données sur plusieurs plateformes afin de répondre à vos objectifs de découverte et aux exigences du Lacuna Fund énumérées ci-dessus. L'utilisation d'un dépôt populaire pour publier l'ensemble de données vous donnera un moyen de préserver l'ensemble de données au-delà de la durée de vie des projets spécifiques. Les plateformes d'hébergement potentielles pour les bénéficiaires de subventions incluent, mais ne sont pas limitées à :

Domaine	Sites d'hébergement
Général	Zenodo - répond à toutes les conditions d'hébergement Registre des données ouvertes de l'AWS DataCite Dataverse Figshare Kaggle
Observations de la Terre	Radiant MLHub
PNL	Les référentiels spécifiques aux tâches, tels que :

Domaine	Sites d'hébergement
	• Dépendances universelles pour l'étiquetage des parties du discours (POS) • OPUS pour les corpus parallèles • Voix commune pour les données vocales non étiquetées avec texte source CC0 Hugging Face Référentiels spécifiques aux langues, tels que ELRA
Santé	Vivli Nightingale Open Science Référentiels d'ensembles de données partagés, tels que AIMI, Physionet
Régional (données ouvertes axées sur l'Afrique)	openAFRICA

Annexe B : Modèle de métadonnées

Identifiant persistant :	*DOI url*	e.g. https://doi.org/10.18653/v1/p19-1346
Titre :		
Créateur(s) de l'ensemble de données :		
Résumé du jeu de données :		
Publication(s) :		
Contributeur(s) :		
Type de données :		
Structure du jeu de données :	*Champs de données, fractionnement des données*	Par exemple, formation, validation, échantillons d'essai.

Point de contact :		
Mots clés/Tags :	*mots-clés supplémentaires que l'on peut utiliser pour trouver cet ensemble de données*	
Langue(s) :		
Licence :	CC-BY 4.0	
Taille :		
Annotations :	*D'où proviennent les annotations dans l'ensemble de données ? Processus d'annotation ?*	Par exemple, source collective, généré par des experts, autre.
Relations avec les travaux existants :	*L'ensemble de données contient-il des données originales et/ou a-t-il été étendu à partir d'autres ensembles de données ?*	par exemple, original, étendu
Fréquence de mise à jour prévue :		Par exemple, trimestriellement, annuellement, autre.
Considérations sur l'utilisation des données :	*Impact social de l'ensemble de données, biais, risques et limites*	

Annexe C : Modèles de fiches techniques

Modèle de feuille de données dans différents formats basé sur la publication *Datasheets for Datasets* de Gebru et al.

- Fiche technique sur le latex
- Feuille de données sur le démarquage
- Fiche technique JSON