Guidance for Dataset Hosting and Documentation

Lacuna Fund: Our Voice in Data

February 2022; Updated June 2024

This document provides guidance to grantees related to the hosting, documentation, and licensing of datasets created with Lacuna Fund support. If you have comments, questions, or suggestions please email the secretariat@lacunafund.org.

Accessibility and Hosting

Lacuna Fund's hosting policy is flexible to allow for local ownership of data and to be responsive to envisioned use cases for the data. Pursuant to Lacuna Fund's principles, dataset hosting should be widely accessible and enable socially beneficial use of the data and further development of the open data in collective ways and as a common good

Lacuna Fund requires grantees to host data on a site that meets the following conditions, in order for datasets to be findable and to track usage:

- Assigns a digital object identifier (DOI) for datasets or allows one to be attached as part of the metadata.
- Is indexed by major search engine(s) (e.g., Google Dataset Search or similar tools).
- Is reliable and persistent.
- Quantifies the number of landing page views and downloads for the dataset.
- Collects contact information for dataset downloads in a way that maximizes conversion.

Consider hosting solutions that:

- allow for integration with tools commonly used to process the data
- are already used by communities who might implement envisioned use cases
- facilitate interaction and exchange around the dataset to make it more useful for use-case development
- enable strong sustainability planning or a governance model for the data
- enable collective further enlargement of the data
- collect some information from users who download the data (e.g., intended purpose for use)

Additional information and examples of potential hosting platforms are available in Appendix A.

Dataset Documentation

Lacuna Fund recognizes that clear documentation is critical to ensuring the accessibility and use of the dataset. **Lacuna Fund asks grantees to include the following documentation when datasets are submitted:**

1) Metadata file (see Appendix B for a template)
2) Datasheet (see Appendix C for templates)
3) Digital object identifier (DOI)

The following links provide general and domain-specific standards and resources:

- Standards for metadata – for example, Spatio-Temporal Asset Catalog (STAC) for earth observation-related data.
- Standardized documentation for ML datasets, such as Datasheets for Datasets (Gebru et. al)
  - For sample models, Model Cards for Model Reporting (Mitchell et. al)
  - For NLP, see Data Statements for Natural Language Processing (Bender and Friedman)
- Taxonomies and other ontologies to incorporate that promote interoperability and usability of datasets.

Licensing

Licensing of datasets must be pursuant to Lacuna Fund's Intellectual Property Policy, which states that: "Datasets and related IP developed with grant funds will: (a) be owned by the subgrantee entity; and (b) be openly licensed by the subgrantee entity to maximize the potential for responsible release and downstream use of such intellectual property. Subgrantee will prioritize releasing the intellectual property under a permissive open-source licensing structure, such as Apache 2.0 (https://opensource.org/licenses/Apache-2.0) for any code or other inventions, or CC-BY 4.0 International (https://creativecommons.org/licenses/by/4.0/), or CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/) for any other intellectual property (e.g., creative works that are not code or patentable); provided that, on a case-by-case basis, Technical Advisory Panels may exercise reasonable discretion regarding more restrictive licensing structures in order to protect privacy, prevent harm, or otherwise maximize the potential for responsible release and downstream use."

"Related intellectual property" generally includes, but is not limited to:

- Metadata, datasheets, or other information.
- Guidance included with the dataset to guide its use.
- Sample models or other informational tools released with the data.

Technical Advisory Panels may exercise reasonable discretion regarding more restrictive licensing structures in order to protect privacy, prevent harm, or otherwise maximize the potential for responsible release and downstream use. **Requests for exceptions to a CC BY 4.0 or CC BY-SA 4.0 International dataset license should be made at the proposal stage**. If an exception is granted, it will be embedded in

your contract. If there have been material changes in the circumstances of your project that would lead to a need for different licensing, please notify the Secretariat as soon as possible.

Appendix A: Hosting Options and Additional Guidance

Please communicate with the organization that will be hosting your dataset at the beginning of your project to understand timelines for dataset release and potential costs to be shared, if any, especially for hosting sites that manually curate and upload data. Information on the process you will undertake to publish your data will be requested in the midterm report.

Lacuna Fund grants require that your dataset be publicly and openly available by the end of the period of performance. If presentation at a conference or publication of academic work may cause a delay in making your dataset openly available, consider a preprint server or a dedicated site before the final hosting for the dataset is established that includes clear information about licensing and data governance.

Publication in an academic journal is NOT required. However, if you do publish an article related to your dataset, we'd love to hear about it!

The following hosting options meet some or all of the requirements outlined in the main guidance above. It is fine to host your data on multiple platforms to meet your goals for discoverability and Lacuna Fund's requirements listed above. Using a popular repository to publish the dataset will give you a way to preserve the dataset beyond the lifespan of specific projects. Potential hosting platforms for grantees include, but are not limited to:

Domain	Hosting sites
General	Zenodo - meets all hosting conditions AWS Registry of Open Data Data Cite Datadryad Dataverse Figshare Kaggle
Earth Observations	Radiant MLHub Meta-repositories / hubs / catalogs of datasets - it is advised that you advertise your dataset by including references to it in as many meta-repositories of datasets as possible::

Domain	Hosting sites
NLP	Task specific repositories, such as: • Universal Dependencies for Part of Speech (POS) tagging • OPUS for parallel corpora • Common Voice for unlabeled speech data with CC0 source text Hugging Face Language-specific repositories, such as ELRA Meta-repositories / hubs / catalogs of datasets- it is advised that you advertise your dataset by including references to it in as many meta-repositories of datasets as possible: - For African NLP: Lanfrica - connecting all African language resources
Health	Vivli Nightingale Open Science Shared datasets repositories, such as AIMI, Physionet Meta-repositories / hubs / catalogs of datasets - it is advised that you advertise your dataset by including references to it in as many meta-repositories of datasets as possible:
Regional (open data focused on Africa)	openAFRICA

Appendix B: Metadata Template

Persistent Identifier:	*DOI url*	e.g. https://doi.org/10.18653/v1/p19-1346
Title:		
Dataset Creator(s):		
Dataset Summary:	[YOUR SUMMARY OF THE DATASET, it would be nice to also include the following paragraph:]	

	This dataset was originally created with support from Lacuna Fund, the world's first collaborative effort to provide data scientists, researchers, and social entrepreneurs in low- and middle-income contexts globally with the resources they need to produce labeled datasets that address urgent problems in their communities. Lacuna Fund is a funder collaborative that includes The Rockefeller Foundation, Google.org, Canada's International Development Research Centre, the German Federal Ministry for Economic Cooperation and Development (BMZ) with GIZ as implementing agency, Wellcome Trust, Gordon and Betty Moore Foundation, Patrick J. McGovern Foundation, and The Robert Wood Johnson Foundation. See https://lacunafund.org/about/ for more information.	
Publication(s):		
Contributor(s):		
Data Type:		
Dataset Structure:	*Data Fields , Data Splits*	e.g. Training, Validation, Test samples
Point of Contact:		
Keywords/Tags:	*additional keywords one can use to find this dataset* Lacuna	
Language(s):		
License:	CC-BY 4.0	
Size:		

Annotations:	*Where do the annotations in the dataset come from? Annotation process ?*	e.g. Crowdsourced, Expert generated, Other
Relations to existing work:	*Does the dataset contain original data and/or was it extended from other datasets?*	e.g. Original, Extended
Expected Update Frequency:		e.g. Quarterly, Annually, Other
Considerations for Using the Data:	*Social Impact of Dataset, Bias, Risks, and Limitations*	

Appendix C: Datasheet Templates

Datasheet template in various formats based on *Datasheets for Datasets* publication by Gebru et al.

- Latex Datasheet
- Markdown Datasheet
- JSON Datasheet

See also: The Data Cards Playbook: A toolkit for transparency in AI dataset documentation