



Lacuna Fund

Dataset Impact Assessment Report

Assessing Use, FAIR Alignment, and Sustainability of Open Datasets for Machine Learning in Low- and Middle-Income Countries

January 2026

Prepared by Datawise Africa



**Datawise
Africa**

 **Lacuna Fund**
Our voice in data

Table of Contents

1. Executive Summary	3
2. Introduction	5
2.1. Background	5
2.2. Purpose	5
2.3. Objectives	6
3. Impact Assessment Overview	7
4. Methodology.....	9
5. Key Findings	14
5.1. Thematic and Domain Distribution	14
5.2. Dataset Availability and Hosting	19
5.3. Alignment with FAIR Principles.....	22
5.4. Evidence of Dataset Usage and Reuse	24
5.5. Sustainability Plans.....	28
6. Challenges and Gaps.....	31
(a) Assessment-Level Challenges	31
(b) Dataset Development-Level Challenges	31
7. Recommendations	34
7.1. Summary	35
8. Conclusion.....	36

1. Executive Summary

This report presents the findings of a detailed impact assessment of 47 datasets funded by Lacuna Fund, a collaboration between The Rockefeller Foundation, Google.org, Canada's International Research Development Centre (IDRC), Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH (GIZ), Wellcome Trust, Patrick J. McGovern Foundation, Gordon & Betty Moore Foundation, and Robert Wood Johnson Foundation.

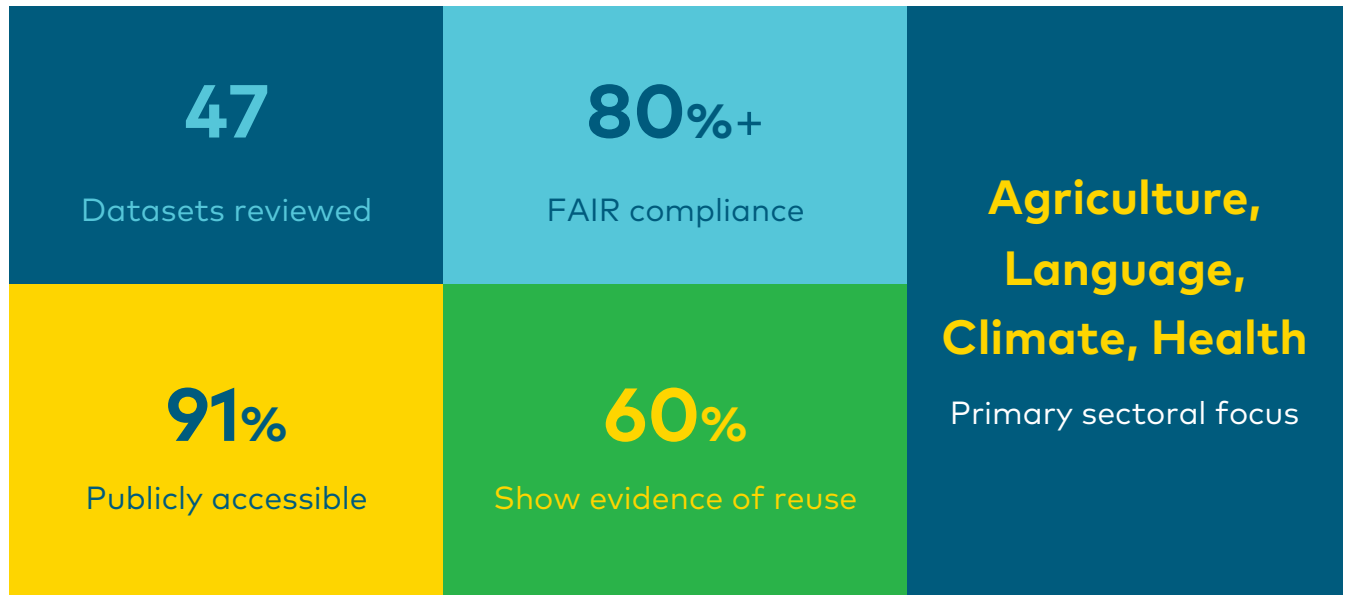
These datasets developed to support AI and machine learning applications in low- and middle-income contexts were evaluated for their adherence to FAIR data principles, usability, citation, sustainability, and real-world impact.

The assessment involved data collection through a structured Google Form, supplemented by desktop research, usage tracking, and stakeholder follow-ups.

Evidence gathered across multiple domains demonstrates that Lacuna Fund-supported datasets have been actively reused in research, model development, and policy-relevant applications. This reuse spans agriculture, language technology, health, and climate, with visible outputs including academic publications, fine-tuned machine learning models, benchmarking studies, and publicly available repositories.

Overall, the findings highlight the strong FAIR compliance and high accessibility of these open datasets, underscoring their growing relevance for unlocking innovation across research, policy, and applied sectors. The report also identifies persistent challenges related to licensing clarity, documentation quality, discoverability, and long-term maintenance.

Recommendations include clearer licensing frameworks, improved metadata standards, more robust hosting plans, and stronger engagement with user communities. The report concludes with proposed actions to further strengthen the sustainability, accessibility, and long-term value of Lacuna Fund-supported datasets.



2. Introduction

2.1. Background

Lacuna Fund was established to address the global gap in quality training and evaluation datasets. By funding locally relevant datasets, it aims to empower data scientists, researchers, and institutions to build AI and ML solutions that are contextually appropriate, equitable, and impactful.

To date, a number of datasets have been published under this initiative, spanning sectors such as agriculture, healthcare, natural language processing, and climate. Lacuna Fund Steering Committee commissioned this assessment to understand whether and how these datasets are being used, the extent to which they are benefiting people, and to evaluate their discoverability, accessibility, reuse, and long-term maintenance.

2.2. Purpose

Purpose	Description
Evaluate Use	Assess if and how datasets are being used
Understand Impact	Gauge influence on research, policy, innovation
Identify Gaps	Surface technical/institutional sustainability challenges
Propose Recommendations	Offer ways to improve future efforts

This assessment also supports Lacuna Fund's commitment to learning, transparency, and continuous improvement by providing evidence-based insights into the lifecycle and usage of the datasets it funds.

2.3. Objectives

The specific objectives of the project were:

Objective	Focus
Usage Tracking	To document how, where, and by whom the datasets are being used.
FAIR Evaluation	To assess datasets against the FAIR principles of Findability, Accessibility, Interoperability, and Reusability.
Impact Stories	To collect qualitative evidence of the datasets' impact on communities, institutions, and research.
Sustainability Check	To determine whether datasets have ongoing maintenance or institutional support.
Recommendations	To provide concrete guidance for enhancing dataset quality, sustainability, and uptake.

3. Impact Assessment Overview

Lacuna Fund Dataset Impact Assessment reviewed 47 machine learning-ready datasets developed to address data gaps in low- and middle-income countries. These span sectors such as agriculture, health, language, and climate, with most created by organisations based in Sub-Saharan Africa, reflecting Lacuna Fund's initial Pan African focus and later expansion to South Asia and Latin America.



The assessment aimed to evaluate:

- The **accessibility, usability, and potential impact** of these datasets
- Their contribution to local innovation and equitable AI development

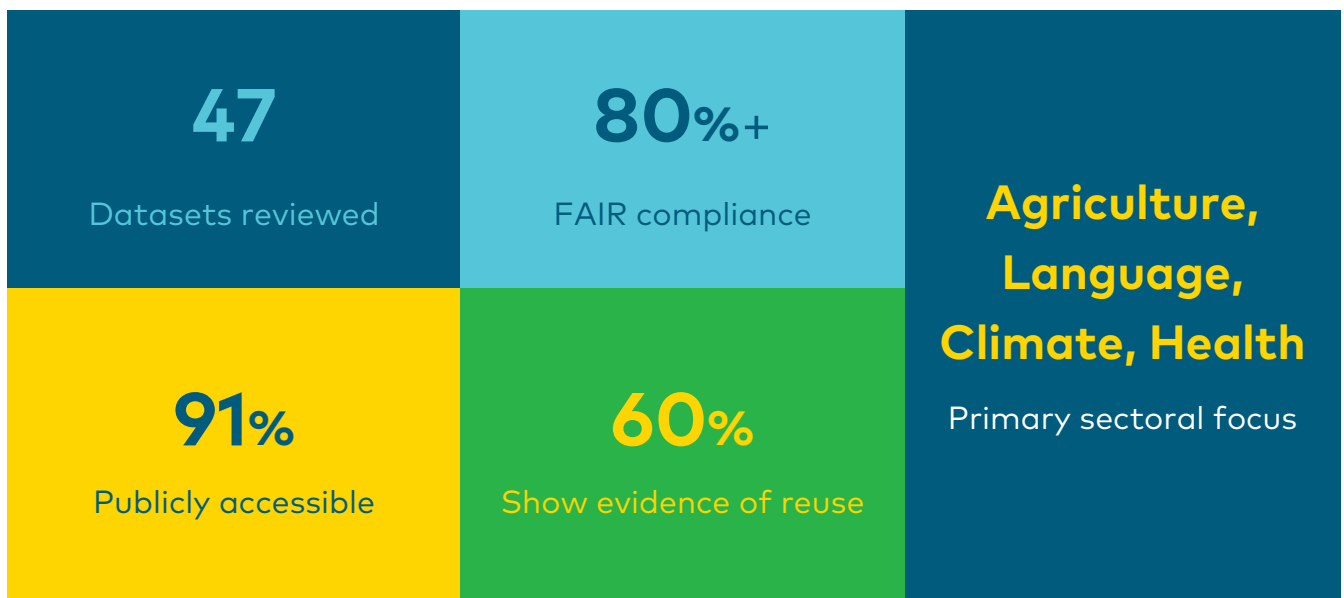
A **mixed-methods approach** was employed, combining:

- Desk reviews and research
- Structured classification
- Quantitative scoring

Six key dimensions were assessed, as outlined below:

1. Alignment with FAIR principles
2. Public availability
3. Target user communities
4. Sustainability planning
5. Evidence of reuse and citation
6. Impact (Popularity and Real-world use)

The results show strong performance in FAIR compliance, with nearly 80% of datasets aligned, and 91% publicly accessible. Agriculture and language datasets accounted for the largest share of published outputs. A growing number of datasets also showed evidence of reuse in tools, publications, or decision-making contexts.



4. Methodology

The assessment employed a mixed-methods framework, combining structured data collection from grantees with digital footprint analysis across open data and AI platforms. The goal was to understand not only whether datasets had been reused, but how, by whom, and to what effect.

1. Dataset Mapping and Classification

All 47 datasets were mapped and categorized into four thematic domains:

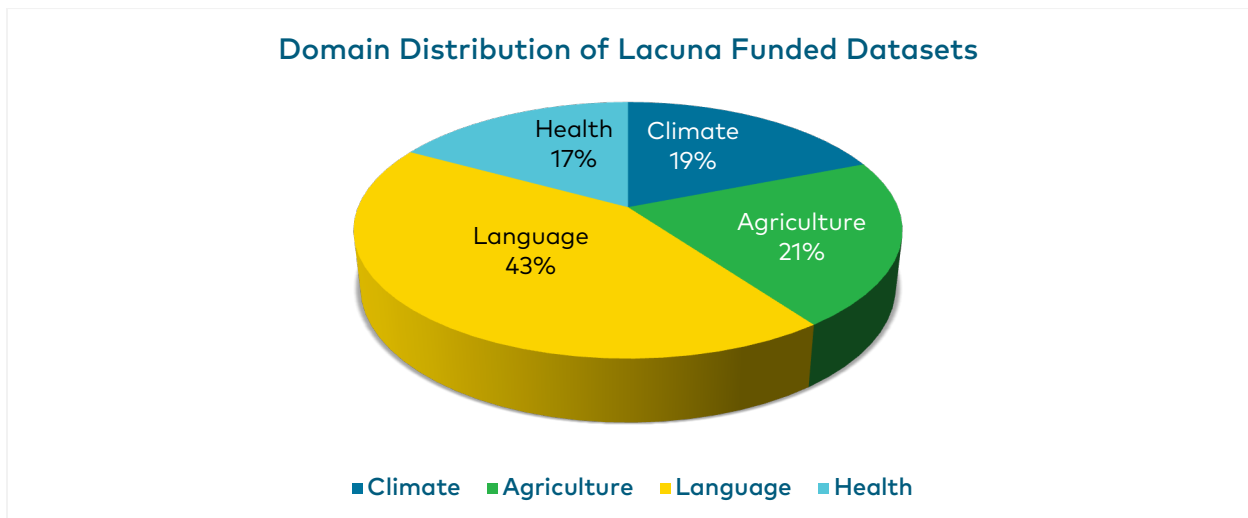


Figure 1: Pie-chart showing domain distribution of datasets

A centralised **Dataset Impact Tracker** was developed using Google Sheets to systematically capture and compare key metadata for all Lacuna Fund-supported datasets. The tracker enabled both quantitative and qualitative analysis across domains, ensuring that patterns in dataset availability, reuse, and FAIR compliance could be assessed consistently.

Each dataset entry in the tracker recorded the following information:

- **Project Title and Owner(s):** The name of the funded project and grantees responsible for dataset development. This allowed tracking of contributions across institutions and sectors.
- **Thematic Area / Domain:** Datasets were classified into four primary domains: Language (NLP), Agriculture, Climate, and Health, enabling domain-level analysis of distribution, accessibility, and reuse trends.
- **Hosting Platform:** The repository or platform where the dataset was maintained, including GitHub, Hugging Face, Zenodo, or institutional repositories. Hosting information was critical for evaluating dataset stability and discoverability.
- **Licensing Type:** Datasets were tagged with license information (e.g., CC-BY 4.0, CC0, or restricted), which informed the potential for reuse and integration into AI pipelines or policy applications.
- **Accessibility Status:** Each dataset was marked as live, in progress, restricted, or unavailable. This allowed rapid identification of datasets with public access versus those facing barriers due to licensing, privacy, or incomplete hosting.
- **FAIR Compliance Metrics:** Basic FAIR indicators (Findable, Accessible, Interoperable, Reusable) were captured to assess adherence to best practices in open and reusable data management.

The tracker was actively used to:

- Guide targeted outreach to dataset owners and grantees for follow-ups or clarifications.
- Identify gaps in dataset documentation, hosting stability, and metadata completeness.
- Analyze patterns in dataset reuse, citation, and integration into research or AI applications.

- Inform reporting on domain-level trends, FAIR compliance, sustainability plans and real-world impact.

Overall, the Dataset Impact Tracker served as the backbone for Lacuna Fund impact assessment, enabling a **structured, evidence-based overview** of dataset availability, FAIR compliance, and reuse patterns across the portfolio.

Dataset Tracker

Lacuna Fund and GIZ-supported Datasets														
CORE METADATA					PROJECT DETAILS					HOSTING AND ACCESS				
Grantee ID	GIZ Funded?	Applicant First Name	Applicant Last Name	Applicant Email Address	Project Title	RFP (Note: For RFP, we are yet to receive any reports since 2020 AG)	Latest completed report	Public URL	Downloadable (Yes/No)	Hosting Platform	Hosting Status	Metadata Available	FAIR: Findable	FAIR: Accessible
AGRICULTURE														
Grantee ID	GIZ Funded?	Applicant First Name	Applicant Last Name	Applicant Email Address	Project Title	RFP (Note: For RFP, we are yet to receive any reports since 2020 AG)	Latest completed report	Public URL	Downloadable (Yes/No)	Hosting Platform	Hosting Status	Metadata Available	FAIR: Findable	FAIR: Accessible
3		Catherine	Nakubutu	cnakubutu@gmail.com	Chapman's Informing Decision-making			https://datacommons.org/dataset/2020-03-01-01	Yes	Zenodo	Live	No	Y	Y
6		Colin	Lee	colin@leed.ac.uk	High accuracy location and field sensor data for	2020 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Google Drive	Live	No	Y	Y
4		Dirine	Dirine Eweon	dirine@univ-lyon1.fr	Machine Learning Systems for Crop Pest and Disease	2020 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Website/GitHub	Live	Yes	Y	Y
2		Joyce Nakubutu	Nakubutu	joyce@univ-lyon1.fr	Machine Learning Systems for Crop Pest and Disease	2020 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Website/GitHub	Live	Yes	Y	Y
7		Lily	Wadhwa	lilywadhwa@gmail.com	Open air environmental image data	2020 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Dataverse	Live	Yes	Y	Y
Yes	29	FALSE	Mutemba	daniel.mutemba@gmail.com	Mobile UGA dataset for enabling rapid	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Zenodo	Live	Yes	Y	Y
33		Colin	Udwin	colin.udwin@univ-lyon1.fr	Human Based Assessment Fish	2020 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Zenodo	Live	Yes	Y	Y
Yes	31	FALSE	World	world@univ-lyon1.fr	A region-wide map of crop yield	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Kaggle	Live	Yes	Y	Y
32		Darlington	Akpo	darlington@univ-lyon1.fr	Open air environmental image data	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Google Drive/Zenodo	Live	Yes	Y	Y
Yes	30	FALSE	Enfer	enfer@univ-lyon1.fr	Human Based Assessment Fish	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Hugging Face	Live	Yes	Y	Y
HEALTH														
Grantee ID	GIZ Funded?	Applicant First Name	Applicant Last Name	Applicant Email Address	Project Title	RFP (Note: For RFP, we are yet to receive any reports since 2020 AG)	Latest completed report	Public URL	Downloadable (Yes/No)	Hosting Platform	Hosting Status	Metadata Available	FAIR: Findable	FAIR: Accessible
Yes	13	FALSE	Bikem	bikem@univ-lyon1.fr	Smartphone artificial intelligence	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes		In progress	No	N	N
Yes	27	FALSE	Bikem	bikem@univ-lyon1.fr	Smartphone artificial intelligence	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Zenodo	Live	Yes	Y	Y
Yes	69	FALSE	Nathan	nathan@univ-lyon1.fr	Human Based Assessment Fish	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Borealis	Live	Yes	Y	Y
Yes	26	FALSE	Kennedy	kennedy@univ-lyon1.fr	Establishment of Machine Learning Models for	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes	Zenodo	Live	Yes	Y	Y
Yes	18	FALSE	Univ-lyon1	univ-lyon1@univ-lyon1.fr	Human Based Assessment Fish	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes		Unavailable	No	N	N
Yes	21	FALSE	Michael	michael@univ-lyon1.fr	Human Based Assessment Fish	2021 AG	Final	https://datacommons.org/dataset/2020-03-01-01	Yes			No	N	N

2. Grantee Engagement via Structured Form

A detailed **Google Form** was designed and distributed to all dataset grantees.

The form focused on gathering:

- Known cases of dataset reuse or integration
- Links to applications, tools, or platforms using the dataset
- Community feedback or testimonials (where available)
- Licensing and hosting details, including any changes since funding
- Challenges faced in ensuring long-term access and reuse

Survey responses were validated and supplemented with manual review of links, tools, and citations provided. Non-respondents were followed up via direct email.

3. Digital Footprint and Citation Tracking

To supplement self-reported data, a thorough public platform scan was conducted using the following tools:

- **GitHub:** To find forks, stars, or projects that reused Lacuna datasets in codebases.
- **Hugging Face:** To identify language datasets used in pre-trained models, fine-tuned transformers, and open-source pipelines.
- **Zenodo, Kaggle,** and institutional repositories: To detect downloads, reuse in competitions, or dataset integration into pipelines.
- **General web search:** To identify media articles, project reports, NGO tools, blog posts, and public dashboards using Lacuna data.

Each identified use case was reviewed and tagged back to its respective dataset in the Tracker.

4. FAIR and Licensing Contextual Review

Although the primary lens of this project was impact, elements of the FAIR Principles (Findable, Accessible, Interoperable, Reusable) were applied to understand barriers to uptake. Special attention was given to:

Focus Area	What Was Assessed
Licensing Clarity	Availability and clarity of dataset licenses (e.g., CC-BY 4.0)
Download Stability	Stability of download links
Metadata & Guides	Quality and availability of metadata or user guides
Hosting Reliability	Whether hosted on institutional repositories or personal sites

Datasets with clear licenses and stable, well-documented hosting environments were found to have higher likelihood of reuse.

5. Desk Research (Grantee Reports)

As part of the desk research, the team conducted a structured review of grantee progress(midterm) and final reports. These reports were analyzed to extract qualitative insights on dataset objectives, development processes, intended use cases, challenges encountered, sustainability plans, and reported evidence of reuse or impact.

Findings from the grantee reports were triangulated with survey responses, digital footprint analysis, and dataset repository reviews to validate claims and enrich the overall impact assessment.

5. Key Findings

This section synthesizes key insights derived from the dataset impact tracker, structured responses from grantees, grantee reports and digital footprint analysis. The findings provide an overview of how Lacuna Fund-supported datasets are performing in terms of availability, discoverability, reusability, and sustainability, with a particular emphasis on adherence to the FAIR data principles.

5.1. Thematic and Domain Distribution

Analysis of Lacuna Fund dataset portfolio reveals a diverse spread across four primary thematic domains: **Language (NLP)**, **Agriculture**, **Climate**, and **Health**. This distribution is intentional: the global, multi-stakeholder steering committee strategically prioritized these domains based on their potential for high impact in low- and middle-income country contexts and it underscores Lacuna Fund's commitment to supporting data creation efforts that bridge regional data gaps and enable locally relevant AI and data-driven research.

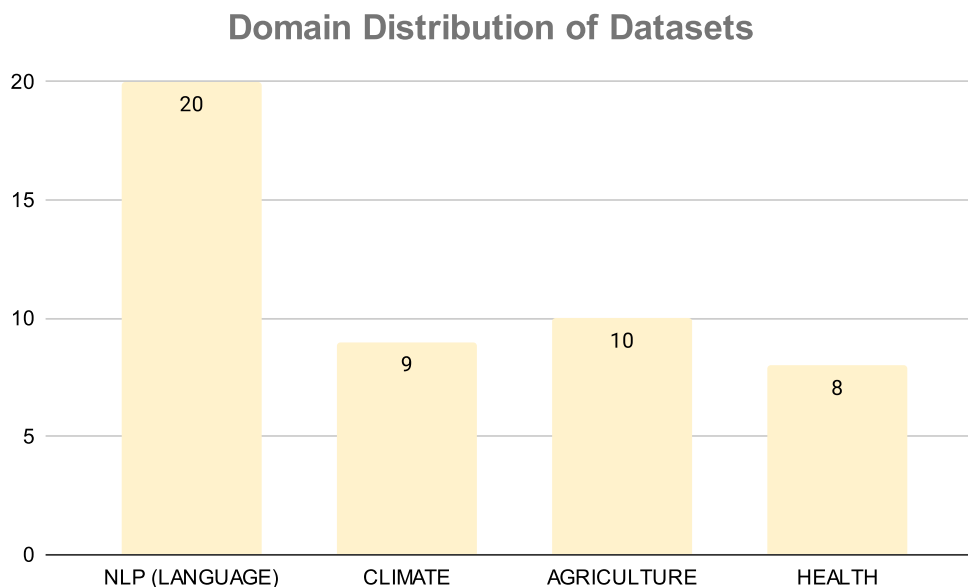


Figure 2: Column chart showing domain distribution

Analysis across the four domains highlights both sector-specific strengths and shared challenges that shape data generation and reuse dynamics within Lacuna portfolio.

1. Language (NLP)

Language datasets form the backbone of Lacuna Fund portfolio, reflecting a concentrated effort to build foundational resources for low-resource languages across Africa, South Asia, and Latin America. These projects collectively address long-standing linguistic inequities that have limited the inclusion of African and Indigenous languages in global AI systems.

The datasets support diverse applications, including machine translation, speech recognition, sentiment analysis, and hate speech detection. Most are hosted on open repositories such as **Hugging Face**, **Zenodo**, and **Kaggle**, which enhances their discoverability, reusability, and integration into downstream tasks.

Examples include **MasakhaNER 2.0**, a corpus for named entity recognition across 20 African languages; **NaijaVoice**, a multilingual speech–text dataset covering Igbo, Hausa, and Yoruba; and the **Bayelemabaga** parallel corpus supporting Bambara–French translation in Mali. Such datasets are now being incorporated into multilingual language models and regional NLP benchmarks, demonstrating strong uptake and tangible contributions to AI inclusivity.

2. Agriculture

Agriculture remains a pivotal thematic domain, directly linked to global food security and climate adaptation agendas. Funded projects have produced datasets that enable precision agriculture, crop monitoring, and smallholder productivity analysis across diverse geographies.

Several initiatives integrate satellite imagery, soil composition data, pest and disease surveillance records, and yield prediction models to support data-driven decision-making. Notable examples include the **Maize Plot Location and Yield dataset in East Africa**, which combines weather, soil, and geospatial data for

yield estimation; and the **CropHarvest dataset**, which informs agricultural development, early warning systems, and trade across Sub-Saharan Africa. Other projects, such as the **Mobile ELISA Livestock Dataset** and the **Sensor-Based Aquaponics Fish Pond Dataset**, extend agricultural data applications into livestock management and aquaculture innovation.

While these datasets are rich in domain content, their reuse is sometimes constrained by incomplete metadata, inconsistent documentation, and limited standardization across repositories. Nevertheless, their long-term analytical value is evident particularly in advancing climate-smart agriculture, nutritional resilience, and resource-efficient farming systems.

3. Climate

Climate-focused datasets represent a growing and strategically important component of Lacuna portfolio. Projects span regions such as Colombia, Madagascar, Tanzania, Ecuador, the Philippines, and Mauritius, addressing themes like solar energy, climate-sensitive diseases, perinatal health, and environmental exposure. These datasets often integrate information from multiple sources including remote sensing, meteorological, and health data to support cross-sectoral analysis.

A recurring challenge identified is **data harmonization**, particularly aligning differences in spatial resolution, temporal coverage, and data collection methodologies. Despite these complexities, the datasets significantly enhance regional capacity for climate-health modeling, renewable energy forecasting, and environmental risk assessment.

4. Health

Health datasets form a vital and high-impact component of Lacuna portfolio, addressing pressing public health challenges through the application of artificial intelligence and machine learning. The projects span diagnostics, disease surveillance, and digital health innovation across multiple African and Latin American countries.

Examples include the **AI-assisted Smartphone Microscopy for Automatic Detection of Diarrhea-Causing Parasites**, which enhances point-of-care diagnostics in low-resource settings, and the Smartphone AI Platform for Paper Record Digitization, which supports the transition from manual to digital health records.

Other projects such as the **Smart Discharges for Mom & Baby initiative** improve maternal and neonatal health outcomes through predictive risk modeling, while the **Africa-BraTS and Tuberculosis Chest X-ray Dataset** projects expand global biomedical imaging resources to better represent African populations.

Additional datasets, including those focused on rabies surveillance, digital pap smear image classification, and childhood malnutrition reduction in Chile, further demonstrate the portfolio's breadth in tackling both infectious and non-communicable diseases.

Health data projects often operate within stricter governance environments due to privacy, consent, and ethical requirements, which limit the extent of open publication. Nonetheless, they provide essential foundations for improving evidence-based health planning and have piloted innovative anonymization and secure data-sharing frameworks that can be scaled to regional systems.

Domain	Representative Datasets / Projects	Primary Focus / Applications	Evidence of Reuse or Impact
Agriculture	Maize Plot Location and Yield Dataset (East Africa); Crop Pest and Disease Diagnosis Dataset	Precision agriculture, crop monitoring, pest detection, yield estimation, smallholder productivity	Widely reused in ML models and EO research; cited in 30+ papers; integrated into agricultural decision-support and trade systems

Health	Smart Discharges for Mom & Baby; Tuberculosis Chest X-ray Dataset for Africa	Disease diagnosis, maternal and child health, public health surveillance, clinical AI model development	Used for AI model training and testing; cited in medical AI studies; enables diagnostic innovation in low-resource settings
Language (NLP)	MasakhaNER 2.0 (NER for 20 African languages); NaijaVoice (Igbo, Hausa, Yoruba speech corpus)	Machine translation, speech recognition, hate speech detection, multilingual model adaptation	High reuse through LLM fine-tuning and benchmarks; cited in ACL/EMNLP papers; integrated into multilingual NLP research pipelines
Climate	Labelled Solar Panel Data for Madagascar; INSPIRE Network (Africa)	Climate-health linkages, renewable energy, urban resilience, predictive modeling	Applied in predictive ML models and climate-health studies; cited in peer-reviewed research; supports policy and resilience planning

5.1.1. Cross-Domain Observations

Cross-cutting analysis reveals that Language datasets exhibit the highest discoverability and reuse, driven by open hosting platforms and global NLP community engagement. Agriculture and Climate datasets are deeply interconnected, often co-applied in modeling crop yields under varying environmental conditions, while Health datasets stand out for their societal importance but face notable restrictions related to privacy and institutional data ownership.

The portfolio as a whole demonstrates substantial progress toward FAIR-aligned practices, particularly in ensuring data accessibility and interoperability in underrepresented contexts. However, persistent challenges across all

domains remain, including metadata incompleteness, inconsistent documentation, and limited cross-platform discoverability.

In summary, thematic and domain distribution analysis confirms that Lacuna Fund has successfully catalyzed data creation across diverse scientific and policy areas. The dominance of language datasets signifies rapid progress in linguistic inclusion for AI, while agriculture and climate datasets strengthen pathways toward sustainable development and environmental resilience. Health datasets, though fewer and more constrained by governance considerations, remain indispensable for national planning and regional health analytics. Together, these domains underscore the Fund's cross-sectoral impact in building the foundation for equitable, data-driven innovation in the Global South.

5.2. Dataset Availability and Hosting

Analysis of the hosting status across Lacuna Fund-supported datasets indicates that the majority are publicly accessible and maintained on stable platforms. Out of 47 datasets assessed, **43 datasets (91.49%) are hosted on reliable platforms** such as GitHub, Hugging Face, and institutional repositories, providing persistent access and enhancing discoverability and reuse potential.

These platforms provide persistent access links, version control, and standardized documentation, supporting long-term discoverability and reuse.

For the purposes of this assessment, dataset availability was classified using the following categories:

- **Live:** The dataset is publicly accessible via a stable hosting platform, with an active download link and sufficient documentation to support reuse.
- **In Progress:** The dataset is under development or curation, with partial outputs available or hosting planned but not yet fully accessible.
- **Restricted:** Access to the dataset is limited due to licensing, ethical, privacy, or institutional constraints.

- **Unavailable:** The dataset could not be accessed at the time of assessment due to broken links, missing repositories, or unresolved access barriers.

A small subset of datasets (4 out of 47, or 8.51%) relies on less stable hosting arrangements, including Google Drive or personal websites. Such platforms pose risks of broken URLs, limited access, and poor sustainability, highlighting the need for migration to more robust repositories.

At the **domain level**, the following patterns emerge:

- **Language (NLP):** All 20 datasets are either live or in progress, with 19 datasets actively hosted on accessible platforms and 1 dataset in progress. No language datasets were restricted or unavailable, reflecting a strong emphasis on discoverability and reuse in this domain.
- **Agriculture:** All 10 datasets are live and publicly accessible, demonstrating solid hosting practices within this sector.
- **Climate:** Of the 9 climate datasets, 6 are live, 2 are restricted, and 1 is unavailable. The restricted and unavailable datasets suggest that licensing limitations or incomplete hosting may affect accessibility.
- **Health:** Health datasets show more variability in accessibility, with 6 datasets live, 1 in progress, and 2 unavailable. These limitations are likely due to privacy, ethical, or institutional constraints associated with sensitive health data.

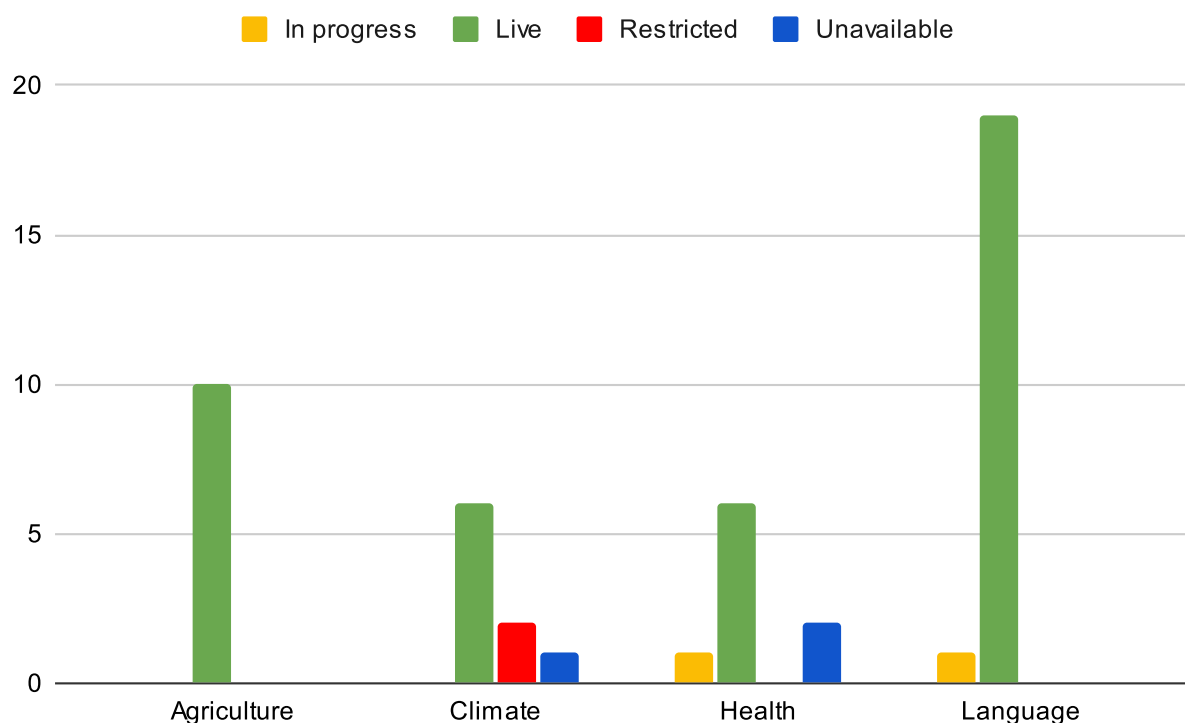


Figure 3: Clustered column chart showing dataset availability per domain

Overall, this domain-level breakdown highlights that **Language and Agriculture datasets exhibit the highest public accessibility**, whereas **Climate and Health datasets face more constraints**, reflecting domain-specific challenges in hosting and sharing.

Overall, the analysis shows that most Lacuna Fund-supported datasets are publicly accessible and hosted on stable platforms, which supports discoverability and reuse. This outcome reflects deliberate guidance from Lacuna Fund and its partners, who developed dataset documentation and hosting guidelines to help grantees identify and use stable, long-term repository platforms.

However, a small portion of datasets remain on less reliable hosting solutions, emphasizing the need for standardized repository practices. These findings provide a foundation for examining FAIR principles, usage patterns, and sustainability considerations in the following section.

5.3. Alignment with FAIR Principles

The majority of assessed datasets demonstrate strong alignment with the FAIR principles, i.e. Findability, Accessibility, Interoperability, and Reusability.

Over 80% of the datasets scored 4 out of 4 across most FAIR categories, indicating that they are well-documented, publicly available, and structured in a way that enables integration and reuse.

AVERAGE FAIR SCORES OF DATASETS PER DOMAIN

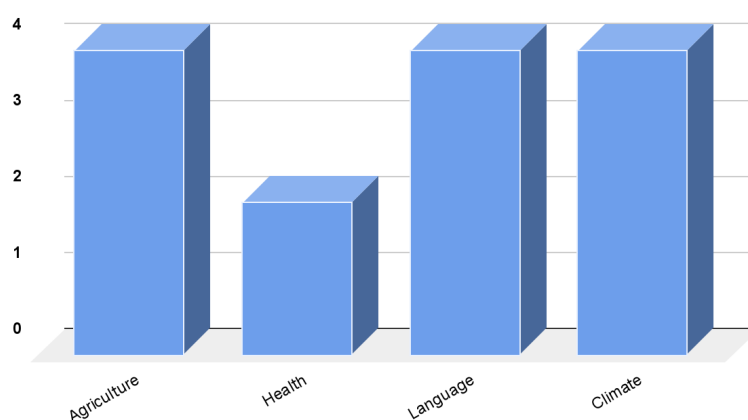


Figure 4: Column chart showing average FAIR scores per domain

5.3.1. Findability

Overall, most datasets demonstrate strong findability, with over 85% assigned persistent identifiers such as DOIs or repository-specific IDs. This is particularly evident in the Language and Agriculture domains, where datasets are hosted on platforms like GitHub, Zenodo, and Hugging Face, ensuring indexing, searchability, and structured metadata. The Makerere NLP datasets for low-resource African languages exemplify best practice: each corpus is fully findable with persistent IDs, clear titles, and descriptive metadata, enabling rapid discovery and integration into multilingual AI models.

However, some Climate and Health datasets lag in this respect. A subset lacks standardized titles, structured keywords, or detailed metadata, while restricted-access datasets common in Health projects remain largely hidden in institutional portals. These limitations reduce visibility in global searches and hinder potential reuse.

5.3.2. Accessibility

Public accessibility is another area of relative strength. Datasets hosted on open repositories generally provide straightforward download options and maintain persistent access. Language and Agriculture datasets, in particular, exhibit high accessibility, with no reported broken links and consistent availability.

Conversely, accessibility challenges persist for certain Health and Climate datasets. Privacy constraints, licensing restrictions, or reliance on unstable platforms such as Google Drive or personal websites reduce continuity of access. In the current portfolio, four datasets fall into this less stable category, underscoring the need for standardized, resilient hosting practices. [Figure 3: Clustered column chart](#) showing dataset availability per domain illustrates the proportion of datasets by domain that are publicly accessible, restricted, or unavailable, providing a visual overview of accessibility patterns.

5.3.3. Interoperability

Interoperability varies considerably across domains. Many language datasets adopt structured formats such as CSV, JSON, or XML, with standardized annotation schemes and UTF-8 encoding, facilitating cross-lingual NLP integration. Similarly, some Agriculture datasets provide harmonized field and crop data suitable for comparative studies and AI modeling pipelines. The Portuguese–Emakhuwa parallel corpus exemplifies a highly interoperable dataset, offering machine-readable aligned text suitable for multilingual NLP tasks.

Health and Climate datasets, however, often use proprietary formats or inconsistent field naming conventions, limiting automated merging and cross-domain analysis. Some climate household survey datasets, for instance, require extensive cleaning before integration, reducing their immediate interoperability.

5.3.4. Reusability

Reusability is strongest where datasets are well-documented and accompanied by clear licensing. Agriculture and Language datasets lead in this area supported by open licenses (CC-BY 4.0, CC0), rich metadata, and example code that enable adoption for research, AI modeling, and policy applications. The NaijaVoice speech corpus illustrates the potential for wide reuse, having been integrated into multiple speech recognition models.

5.4. Evidence of Dataset Usage and Reuse

Evidence gathered across multiple domains demonstrates that Lacuna Fund–supported datasets have been actively reused in research, model development, and policy-relevant applications. This reuse spans sectors including agriculture, language technology, health, and climate, with visible outputs such as academic publications, fine-tuned machine learning models, benchmarking studies, and public repositories.

5.4.1. Popularity by real-world impact

While 47 datasets were analyzed, a selection of representative examples with **verifiable, publicly available evidence** of reuse is highlighted below. These cases exemplify how open, well-documented datasets have catalyzed secondary use and innovation across Africa and beyond.

Domain	Dataset Example	Evidence of Reuse	Impact Insight
Agriculture	CropHarvest (NASA Harvest)	Widely used for training and benchmarking ML models for crop-type classification using Sentinel-2 imagery; over 30 citations in research papers including NeurIPS 2021 Datasets & Benchmarks	High-impact dataset promoting reproducibility, community participation, and AI-driven agricultural monitoring in data-scarce regions.

		Track; reproducible open pipelines on GitHub.	
Language & NLP	KenCorpus: Kenyan Languages Corpus	Publicly available under CC-BY 4.0 license at KenCorpus.co.ke ; reused in university theses, local language technology prototypes, and student AI training. Integrated into Speech-to-Text and Question Answering models; supports Kenyan linguistic inclusion.	Demonstrates academic and cultural reuse, advancing language preservation and innovation in African NLP.
Climate	Tanzania Climate-Sensitive Waterborne Diseases Dataset	Reused in predictive modeling of disease outbreaks under climate variability; referenced in Frontiers in AI (2025); cited in climate-health risk assessment studies; available on Zenodo with metadata and licensing.	Illustrates how open climate-health data informs predictive analytics and evidence-based policymaking for environmental resilience.
Health	African Breast Imaging Dataset for Equitable Cancer Care	Underwent ethics approvals in Uganda and Nigeria; reused in AI hackathons (SPARK AI, PRECISE 2024) with over 270 participants across 28 countries developing breast tumor segmentation models. Hosted community engagement workshops and training for health workers in ultrasound use.	Exemplifies a multi-impact dataset, combining AI research reuse, community training, and clinical skill transfer, strengthening women's health equity and diagnostic innovation in Africa.

Across the 47 Lacuna Fund datasets, several recurring patterns of reuse and impact emerged. A significant number of datasets are being applied in model training and benchmarking, with examples such as *CropHarvest* and *MasakhaNER* now widely integrated into machine learning pipelines for crop classification, named entity recognition, and multilingual language model fine-tuning. This reflects the growing use of locally developed data in shaping global AI research.

Academic citations and publications further affirm this trend, with multiple datasets cited in peer-reviewed papers and presented at international conferences. For instance, the *Climate-Sensitive Waterborne Diseases Dataset (Tanzania)* has been referenced in *Frontiers in Artificial Intelligence (2025)*, while the *African Breast Imaging Dataset* has supported large-scale AI hackathons focused on cancer diagnostics.

Cross-domain reuse is also increasingly visible. Language datasets initially designed for text processing are being adapted for translation, sentiment analysis, and misinformation detection, while agriculture datasets are being integrated into climate and environmental modeling. Similarly, open-source reuse continues to thrive, datasets hosted on platforms such as GitHub, Hugging Face, and Zenodo show active forks, contributions, and derivative works, signaling sustained engagement and capacity building within the data science community.

Policy and research linkages provide further evidence of relevance and uptake. Climate and health datasets are being used in applied research that informs policy analysis, disease surveillance, and agricultural monitoring. These applications demonstrate how open, well-documented data can support evidence-based decision-making across diverse sectors.

The overall picture is clear: Lacuna-funded datasets are not only accessible but actively reused across scientific, community, and policy landscapes. This aligns strongly with the Fund's goal of fostering equitable, high-quality, and locally

grounded data resources that enable innovation and problem-solving in the Global South.

Testimonials from data creators further highlight this expanding impact. As reported by the **International Development Research Centre (IDRC, 2024)**, the *KenCorpus* team shared that *"the dataset has since been downloaded an incredible 250,000 times,"* demonstrating strong discoverability and sustained community engagement with locally developed linguistic resources. Similarly, the *High-Accuracy Maize Plot Location and Yield Dataset* team noted that their project *"improved the usability of the most expansive East African crop yield estimation datasets by correcting the geolocations of the fields,"* thereby enhancing agricultural modeling and supporting evidence-based agricultural decision systems.

Together, these examples illustrate a growing ecosystem of open, interoperable, and impactful data, one that bridges local insight with global research frontiers, advancing Lacuna Fund's vision of equitable AI and data-driven development.

5.4.2. Popularity by Downloads

Download and access metrics across hosting platforms vary considerably, making direct comparisons between datasets difficult. Platforms such as Zenodo and Harvard Dataverse typically report **cumulative download counts** (i.e., total downloads since publication). In contrast, platforms like Hugging Face may present **monthly download or "use" counts** in addition to totals, while GitHub often highlights stars, forks or clones rather than standard download counts.

Despite these measurement differences, several Lacuna Fund-supported datasets stand out for their strong visibility and uptake.

5.5. Sustainability Plans

The assessment revealed a mixed picture of dataset sustainability across the 47 Lacuna-funded projects. While nearly half of the datasets demonstrated clear plans for long-term availability through institutional hosting, continued maintenance, or community engagement, the remainder lacked actionable sustainability strategies.

Four main sustainability pathways emerged from the analysis:

1. Institutional Hosting and Integration

Several datasets have been embedded within academic or research institutions, ensuring continued visibility and infrastructure support. For example, language and legal datasets hosted within university data labs, such as *Kencorpus* at the University of Nairobi, leverage institutional mandates to maintain accessibility.

However, this approach faces **sustainability challenges** such as *bureaucratic delays, limited digital preservation policies, and hosting dependence on internal IT systems*, which pose risks to timely updates and long-term accessibility once project funding ends.

2. Community and Partner Engagement

Some projects sustain their datasets through active collaboration with universities, NGOs, and private-sector partners. This distributed model of stewardship enhances reuse especially in agriculture, climate, and health domains while embedding local capacity and expertise.

Yet, **partnership fragility** remains a key sustainability risk. Many collaborations operate without formal agreements or post-grant funding, making it difficult to ensure continuity when institutional priorities or budgets shift.

3. Open Repository and Licensing Strategy

A substantial number of datasets are hosted on open platforms such as GitHub, Hugging Face, and Zenodo under permissive licenses (CC-BY 4.0 or CC0). This approach promotes discoverability, legal clarity, and broad reuse.

Still, **technical maintenance and metadata decay** emerge as major sustainability challenges. Repositories managed by individual contributors are vulnerable to link rot, missing version control, or outdated documentation, factors that erode dataset reliability and long-term usability.

4. Replication, Documentation, and Capacity Building

Projects that invested in detailed documentation, training workshops, and replication activities demonstrated stronger prospects for sustainability. These practices enable datasets to evolve into open digital public goods reproducible, extensible, and valuable to diverse communities.

However, documentation quality and standardization varied widely, and **replication barriers**, such as missing processing scripts or inconsistent data formats limit the ability of others to maintain or build upon the datasets.

Indicator	Insight / Status
Datasets with a sustainability plan	~50% include clear or partial sustainability strategies.
Common pathways	Institutional hosting, open repositories, partner engagement.
Common gaps	Weak documentation, unclear update or funding plans.
Key risks	Hosting dependency, metadata decay, limited resources.
Emerging good practices	Institutional integration, open licensing, training, and replication.
Recommendation	Embed sustainability planning from project inception.

Overall, sustainability within Lacuna-funded datasets is evolving but uneven. The most resilient datasets are those that are institutionally anchored, openly licensed, and supported by active partnerships and replication efforts.

To ensure long-term impact, future funding cycles should embed sustainability planning from the outset covering hosting, licensing, documentation, replication, and stewardship rather than addressing these issues at the project's end.

6. Challenges and Gaps

Lacuna Dataset Impact Assessment revealed a mix of assessment-level and dataset-level challenges that shaped the depth, accuracy, and sustainability of project outcomes. These issues highlight the broader realities of open data development and stewardship across diverse African research contexts.

(a) Assessment-Level Challenges

The assessment process encountered several practical and structural barriers that limited comprehensive analysis across all 47 datasets.

Some grantee teams were unresponsive or provided incomplete updates, restricting the ability to verify dataset status, hosting, or reuse.

Additionally, platform accessibility and multilingual usability emerged as key challenges. Certain hosting platforms required user registration or presented documentation only in a single language, limiting accessibility and effective evaluation across linguistic communities.

Tracking dataset reach also proved difficult, as several repositories lacked analytics or mechanisms to capture downloads, citations, or derivative works.

Inconsistent metadata and unclear licensing information further reduced discoverability and made it harder to assess compliance with FAIR (Findable, Accessible, Interoperable, Reusable) data standards.

Finally, very few teams had detailed sustainability or maintenance plans, creating uncertainty about long-term dataset viability even for strong initial outputs.

(b) Dataset Development-Level Challenges

During dataset design and implementation, several recurring challenges affected data quality, usability, and continuity.

Technical and infrastructure gaps were common, particularly in data management, documentation, and version control. Some projects relied on temporary storage or personal accounts, raising long-term preservation risks.

Funding limitations also emerged as a key constraint after initial grant periods, most teams lacked resources to maintain or update datasets or support ongoing user engagement.

Where datasets were not integrated into institutional programs, universities, or government systems, long-term hosting and visibility suffered.

Documentation and licensing inconsistencies further constrained reuse, as teams often lacked guidance on open data standards and metadata frameworks.

Finally, ethical review processes and data governance restrictions, particularly in health and language domains, sometimes delayed dataset release or limited public accessibility.

Challenge Area	Description
Grantee Responsiveness	Some teams provided incomplete or no feedback, limiting dataset validation and follow-up.
Platform Accessibility & Multilingual Support	Limited multilingual documentation and complex platform navigation reduced access, interpretation, and assessment of datasets across regions.
Hosting & Tracking Gaps	Many repositories lacked download or reuse tracking tools, making impact measurement difficult.
Metadata & Licensing Issues	Missing or unclear metadata and licenses reduced discoverability and FAIR compliance.

Sustainability Information	Few teams shared maintenance or update plans, leaving dataset longevity uncertain.
Technical Capacity Limitations	Weak data management and documentation practices affected dataset quality and preservation.
Funding Constraints	Lack of post-grant resources hindered updates, hosting, and user support.
Institutional Integration	Datasets not embedded in established institutions were less likely to be sustained or updated.
Documentation Inconsistencies	Variations in documentation and licensing reduced usability and reusability.
Collaboration & Stewardship Gaps	Reliance on individual leads rather than institutional teams led to loss of continuity.
Ethical & Governance Barriers	Data-sharing restrictions, especially in health and language domains, delayed or limited public access.

While the assessment team faced constraints in verifying dataset-level outcomes, the underlying issues point to systemic challenges during dataset development and dissemination. Addressing these, through stronger institutional partnerships, multilingual documentation, and sustained funding models will be essential to ensure Lacuna-funded datasets remain discoverable, usable, and impactful over time.

7. Recommendations

To address the identified challenges and enhance the long-term impact of Lacuna-funded datasets, the following recommendations are proposed:

1. Strengthen Post-Funding Engagement

Establish structured mechanisms for continued engagement with grantees beyond the funding cycle. Supporting sustained dataset curation, usage tracking, and responsiveness during evaluation periods will improve long-term quality and usability. Clear reporting protocols and regularly updated metadata can further enhance discoverability and facilitate impact assessment.

2. Standardize Hosting and Documentation Practices

Promote consistent dataset hosting on trusted open repositories and ensure multilingual documentation to improve accessibility. Standardized metadata formats, version control, and open licensing will strengthen interoperability and long-term preservation.

3. Embed Sustainability and Update Plans in Project Design

Require all future projects to include clear sustainability and update strategies outlining custodianship, feedback mechanisms, and funding pathways. Embedding sustainability early ensures datasets remain relevant and accessible beyond the initial grant period.

4. Build Institutional and Technical Capacity

Encourage integration of datasets into institutional or national repositories to guarantee continuity. Provide targeted training on FAIR data management, metadata standards, and repository governance to strengthen stewardship and reduce dependency on individual project leads.

5. Foster a Collaborative Community of Practice

Promote peer exchange among grantees to share lessons on sustainability,

reuse, and governance. Curate and disseminate good practice case studies such as those demonstrating strong institutional hosting or cross-sector partnerships to build collective capacity and reinforce sustainable open data ecosystems.

7.1. Summary

Embedding sustainability, standardization, and structured post-funding engagement within future grant cycles will ensure that Lacuna-funded datasets remain discoverable, reusable, and impactful over time. Strengthening institutional hosting, promoting collaboration, and aligning with FAIR data principles will position Lacuna Fund as a model for long-term data stewardship and equitable access across Africa and the Global South.

8. Conclusion

This impact assessment reaffirms that Lacuna Fund-supported datasets are making tangible progress toward building open, inclusive, and AI-ready data ecosystems in the Global South. Through 47 funded projects across health, language, agriculture, and climate domains, the initiative has strengthened regional data capacities and fostered collaboration between researchers, governments, and communities.

Many datasets demonstrate high FAIR alignment, public availability, and growing evidence of reuse ranging from citations in peer-reviewed publications to integration into model training, benchmarking, and decision-support systems. These outcomes underscore the Fund's catalytic role in enabling local data ownership and promoting equitable participation in global AI research.

Nevertheless, challenges persist. Gaps in documentation, hosting stability, and sustainability planning continue to limit long-term accessibility and reuse. Strengthening institutional stewardship, embedding sustainability mechanisms at project inception, and maintaining post-grant engagement will be crucial to consolidating these gains.

As the ecosystem matures, the findings from this assessment provide a foundation for strategic reinvestment, continuous learning, and coordinated monitoring across future Lacuna-supported initiatives. Ensuring that these datasets remain living, accessible, and trusted public goods will be key to realizing their full impact on research, innovation, and policy in the years ahead.

Acknowledgements

This assessment report was funded by the **Wellcome Trust**. The analysis, findings, and conclusions presented in this report are those of the authors and do not necessarily reflect the views of the funder.



Prepared by:

Datawise Africa

Website: www.datawiseafrica.com

Email: info@datawiseafrica.com



**Datawise
Africa**

 **Lacuna Fund**
Our voice in data