



DEVELOPMENT
DATA PARTNERSHIP



Unlocking Data Silos for AI Applications

Supported by: **Gates Foundation**



Program Objectives

A scalable, rights-based approach to unlocking proprietary datasets for building public-good AI applications

This Gates Foundation–funded initiative aims to **develop a scalable, replicable model for unlocking proprietary data for use in AI applications**. The first part supports a framework for building national low-resource language libraries derived from media company and government content, and the second supports facilitating scaled access to donor grantee datasets.

Current efforts to gather low-resource language data rely on scraped web content and isolated field recordings. While useful, these sources are fragmented and insufficient for training high-performing models. As a result, speakers of these languages are often left behind. This project addresses the gap by using the Development Data Partnership's infrastructure to license and prepare high-value content—such as news, books, radio, and surveys—under clear consent, privacy protections, and non-commercial terms. The resulting rights-based libraries will be transferred to national stewards to support continued corpus growth, AI training, and inclusive digital transformation rooted in local languages and priorities.



Similarly, donor grantee datasets are often siloed under varying legal terms and organizational policies. This initiative seeks to harmonize governance and technical frameworks to streamline responsible access



The Development Data Partnership

Introduction





Why Data Partnership?

- Public policy and provision of public infrastructure and services are heavily dependent on data – higher quality, timely data translates into more effective sector and program prioritization, design, implementation, monitoring, and evaluation.
- Increasingly, the private sector is generating data that could be used to complement traditional public-sector data collection methods. Through public-private collaboration, in addition to generating more timely and relevant data for decision making, entirely new public good use cases could be discovered and implemented.



Development Data Partnership Goals



Coordinate and aggregate private sector data demand on behalf of public actors



Link public sector challenges and domain expertise to relevant proprietary data



Increase transparency and accountability for integrating use of proprietary data in public good analytics



Reduce international organization duplication of effort and facilitate collaboration on solutions



Significantly reduce the transaction costs associated with data sharing



Increase the capacity of the public sector to procure and use proprietary data and platforms for better decision making



Our Partners | Companies





Our Partners | International Organizations



Partnership Frameworks

The Partnership supports international development through data sharing and data science collaboration between companies and international organizations, leveraging three frameworks:



Legal



Governance



Technical





How The Partnership Works

1

Data Partnership

Data License Agreements are signed between the data partner and each participating organizations (e.g. World Bank, IMF, IDB, OECD, UNDP)



2

Project Proposals

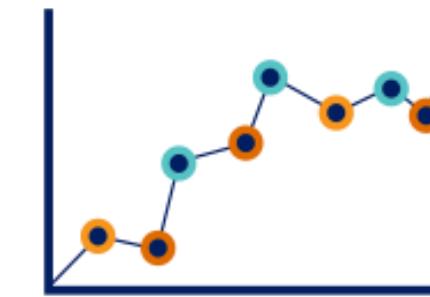
Data partner uses the Partnership Portal to evaluate all proposals for using their data in a project.



3

Data Management

Upon project approval, data are securely and responsibly managed on behalf of all staff through the Partnership IT architecture and procedures.



4

Data Goods

Data partners are kept updated through the Partnership Portal on derived data products and produced code.





The Partnership Today



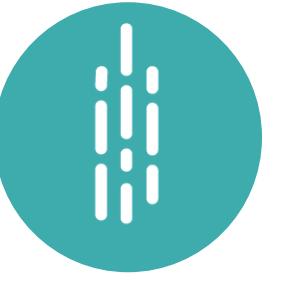
Scalable Partnership Model

Each year, international organization participation continues to grow, with 11 partners today.



Over 400 Development Projects

Teams from across membership organizations are leveraging data provided by the private sector to address global challenges and the Sustainable Development Goals.



Growing Data Partners

30+ active data-sharing agreements signed with private companies to bridge private sector data and public sector needs.



Impactful Results

Insights and data goods generated by the projects are informing public policy and government decision-making across different countries and regions.



Learn About Our Projects

Leveraging LinkedIn Data to Understand the Green and Digital Transformations in the Labor Market and the Future of Work

The economy and the labor market have evolved due to digital and green transitions. New technologies and new forms of work stemming from digitalization, climate change, and ongoing efforts to shift to a low-carbon economy have been changing jobs and skills at a rapid pace.

[Learn More](#)

Advancing Women's Empowerment Through Data

[Case Study](#)

Data plays a fundamental role in identifying the challenges women are facing and in allowing policymakers and businesses to make informed decisions that improve women's well-being and access to economic opportunities and overall development outcomes.

[Learn More](#)

How to Leverage Data for Better Transport, Digital Connectivity, and Sustainable Development in the Amazon

[Case Study](#) [Google](#) [Mapbox](#) [Meta](#) [Ookla](#)
[Digital Development](#) [Transport](#)

The World Bank Transport team used data from Google, Mapbox, Meta, and Ookla to identify infrastructure gaps in the Amazon areas of Brazil, Colombia and Peru.

[Learn More](#)

Unleashing the Power of Data to Tackle Traffic Congestion and Promote Road Safety

[Case Study](#) [Mapbox](#) [Mapillary](#) [Waze](#)

As the global population is constantly increasing, building modern and sustainable cities to accommodate everyone is crucial. This is enshrined in the 2030 Agenda for Sustainable Development, in which the Sustainable Development Goal 11 aims to make cities more inclusive, safe, resilient and sustainable.

[Learn More](#)

Can private management of African protected areas improve socioeconomic and wildlife outcomes?

[Case Study](#) [AtlasAI](#)
[Inequality and Shared Prosperity](#)

African governments increasingly entrust the management of protected areas to private, nongovernmental organizations, with the hope that the significant resources and technical capacities of private organizations will help realize the potential of these areas.

[Learn More](#)

Leveraging Social Media Data to Map Road Traffic Crashes

[Case Study](#) [X](#)

Challenge Road traffic crashes are among the world's most pressing public health challenges. Crashes are the leading cause of death for those 5-29 years old and are the 8th leading cause of death considering all ages.

[Learn More](#)

Impact Stories



Component 1

Unlocking Low-Resource
Language Data





Unlocking Low-Resource Language Data Silos

Component 1 | Low-Resource Language Data

Component Objective

To create a scalable, replicable model for building national digital language libraries for training and tuning AI models.

Challenge

~4 billion people speak low-resource languages, which are often overlooked in AI model training and applications. Widely spoken low-resource language content can exist but is often proprietary and locked in silos.

Solution

Leverage Partnership legal, governance, and technical frameworks to unlock data from media companies and government agencies for use in public good AI model training and applications.



Why Secure Licensing to Complement Open Data Initiatives

Ethical Compensation Framework | Current AI training practices create an unfair extraction model where creators provide value without receiving compensation. A secure licensing system establishes fair value exchange where all creators are compensated proportionally to their contribution, regardless of geography.

Data Scarcity Reality | Most low-resource languages have less than 1% of the data available for English models. Open data mandates cannot bridge this gap because the high-quality content doesn't exist in open repositories, and manual collection efforts can be challenging to scale and sustain.

Quality and Safety Benefits | Copyrighted content offers advantages through professional editorial processes and fact-checking that scraped data can lack. This results in higher linguistic accuracy. Using professionally curated content reduces the substantial costs of bias mitigation and content moderation required when scraping unfiltered internet data.

Economic Sustainability | This system creates sustainable revenue streams for publishers and content creators while enabling breakthrough research. Market mechanisms drive investment in underrepresented languages more effectively than charity models, creating a self-sustaining ecosystem where quality content fuels quality research and linguistic diversity thrives.

'Impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says

Pressure grows on artificial intelligence firms over the content used to train their products

<https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai>

OpenAI Strikes a Deal to License News Corp Content

The deal gives OpenAI's chatbots access to new and archived material from The Wall Street Journal, The New York Post, MarketWatch and Barron's, among others.

"...the Wall Street Journal reported the agreement could be worth as much as \$250 million over five years."

<https://www.nytimes.com/2024/05/22/business/media/openai-news-corp-content-deal.html>



Approach: Learning by Doing

Component 1 | Low-Resource Language Data

1

Legal Framework | Curation

The team will engage and source content from media companies and government agencies in a pilot country, leveraging a modified version of the DDP master data license agreement. The team will set up a data review and cleaning pipeline and meta data schema tailored for proprietary low-resource language libraries.

2

Technical Framework | Pre-Processing

The team will develop, document, and implement methods for pre-processing text and audio data received from different sources -- .pdf newspaper scans, household survey audio recordings, .xml newspaper web pages, etc. The team will create a data flow that ensures original IP are not fully disclosed through the library, including methods for chunking and shuffling data, linking chunks to context, and making them accessible via a monitored API.

3

Technical Framework | Library Management

The team will leverage the current Partnership processes for managing ad fulfilling data requests, with the addition of a new externally facing application page and catalog. The team will prepare a model verification tool to ensure original proprietary contents cannot be 'leaked' through the model, as well as other guidance and benchmarking tools for users.

4

Governance Framework | Scaling & Sustaining

Finally, the team will support knowledge transfer to the pilot country government for replicating the library establishment and process, supported by a percentage of funds from third party companies purchasing updates of the library contents. All outputs will be packaged for use in new countries just starting their libraries.



Use Case: Malawi Implementation

Component 1 | Low-Resource Language Data

- Low Resource Language
 - Chichewa spoken by 20+ million people (of which ~25% speak English)
- Government Partners
 - Ministry of Information:
 - Key national language library stakeholder
 - Will use project-derived LLMs to support government e-services program
 - National Statistical Office
 - Providing household survey audio recordings and transcriptions
 - Will use project-derived ASR model to support survey supervision
- Media Partners
 - Organized workshop with 27 media, academic, government, and NGO stakeholders
 - Bi-lateral negotiations with 18 companies, using team-developed legal terms and cost recovery templates



Curation

Stakeholder Engagement

Objective: Efficiently engage with a wide range of content stakeholders, ensuring transparent communication about data provenance and planned usage.

Approach: Kick-off by organizing a publicly communicated and government-supported stakeholder workshop with content providers for awareness and inputs. Engage in follow-up negotiations using a "terms sheet" for efficient and transparent communication about data license agreement contents.

In Practice: Media stakeholder workshop held in Malawi in 3/2025, with 27 media stakeholders – newspapers, radio, book publishers, writers' union, copyright association, and universities – as well as senior government leadership from the Ministry of Information and the National Statistical Office. Workshop followed by 18 bi-lateral stakeholder meetings using the **terms sheet**, securing buy-in for content provision.



	<p>Organization can confirm that you have the authority to license <u>these content</u> for the sole purpose of training AI models.</p> <p>Proposed Terms:</p> <ul style="list-style-type: none">- Your organization is requested to only provide content for which you fully control the rights.- In the event the provided content is mixed – that is, it contains a mix of content for which your organization has full rights and does not have full sub-licensing rights (e.g., music, advertisements), your organization will notify the project team so that these contents can be removed from the project.- The content would be licensed with "perpetual" rights, because the trained models will continue to benefit from the content data used to train them, even if the original dataset is deleted and irrecoverable.- The license to use these data for model training would be "irrevocable", that is, it is understood by all parties that once content <u>are</u> used to train a model, the content cannot be 'removed' from the model.	<p>There is also foreign content – e.g., partnerships with foreign channels (e.g., BBC, NHK in Japan, CCTV) – but these aren't in Chichewa.</p> <p>~90% content <u>owned</u> by MPC.</p>
3. Non-Commercial Usage	<p>Project Team Understanding (for Confirmation)</p> <ul style="list-style-type: none">- The project is solely for public benefit. No direct commercial exploitation or resale of your organization's content is permitted under this project. <p>Proposed Terms:</p> <ul style="list-style-type: none">- The content licensing for the project will involve no royalties or ongoing fees. Instead, if certain staff tasks are needed to locate, extract or format the recordings, the project team can agree on a cost recovery lumpsum payment. This lumpsum must be itemized (e.g., staff hours, scanning costs, cloud transfer costs).- Once paid, no further or recurring usage fees for the content would be supported.	Confirmed.
4. Sub-Licensing	<p>Project Team Understanding (for Confirmation)</p> <ul style="list-style-type: none">- The project team may sub-license transcripts with researchers and (academic institutions, international) under the same perpetual, non-exclusive license. <p>Proposed Terms:</p> <ul style="list-style-type: none">- All sub-licensees are bound by the terms sheet about moral rights, personal data protection, and data security.	Confirmed. Noted regarding MBC Digital's post

<p>MBC Digital</p> <p>March 27 · 4</p> <p>#Update</p> <p>World Bank, in collaboration with the Ministry of Information and the National Statistical Office, is set to embark on a project that will develop a National Chichewa Library for training AI models that will increase Chichewa content on internet search engines.</p> <p>As part of the initiative, the World Bank organised a day-long workshop in Blantyre for media personnel and book publishers to solicit views on how this can be achieved.</p> <p>Speaking during the event, World Bank's Data Lab Programme Manager, Holly Krambeck, said the library will be used to train AI models, enabling the public to access AI models in Chichewa, thereby increasing access to information in the local language.</p> <p>Publishing Editor for Acin, Alfred Msadala, commended the initiative, saying it will cater for a larger Malawian audience which do not speak or write English.</p> <p>By Simeon Boyce #MBCDigital #Manthu</p>
--





Curation

Legal Framework

Objective: Secure rights to proprietary content from media firms for distribution through a library and for use in training AI models, while minimizing negotiation requirements and protecting both content providers and users.

Approach: Leverage Development Data Partnership Master Data License Agreement (MDLA), which has been signed by more than 40 international organizations and private companies for use in sharing sensitive and proprietary data. Supplement with a new AI Annex, which grants sub-licensing rights, permissions for AI model training, and protections for content providers and users.

In Practice: To ensure fairness and scalability, the new MDLA content has been reviewed and revised pro-bono by Freshfields, a global-leading law firm.



Freshfields Bruckhaus Deringer

Freshfields Bruckhaus Deringer is one of the world's leading law firms, part of London's elite "Magic Circle," with over 275 years of experience advising on complex global deals and disputes.

Their decision to provide **pro bono support** on our AI content licensing agreement brings **world-class legal expertise—normally reserved for high-stakes corporate clients—** directly to our initiative. This **underscores the significance of our project and the trust placed in its global relevance.**



Curation

Meta Data Schema

Objective: Make data as discoverable and usable as possible for AI model training and fine tuning.

Approach: Leverage the World Bank's open-source [Meta Data Editor](#) to create template for an updated meta data schema that meets the unique needs of a text and audio content library for model training.

In Practice: The team is currently developing the schema and template. The team will work with content providers to ensure templates are completed. Over time, some components may become automatable (e.g., context summaries).

Metadata Editor

A multi-standard open-source metadata editor
World Bank, Office of the Chief Statistician

User and practice guide

User Guide in HTML
A searchable version of the user and practice guide, including installation instructions, and step-by-step instructions.

User Guide in PDF
A downloadable version of the user and practice guide.

User Guide in Markdown
Use this link if you want to contribute to the documentation on GitHub, by providing content or suggestions.

Metadata Editor
The Metadata Editor's GitHub repository. Access the application, contribute to the code, submit issues and suggestions.



Pre-Processing

Making Data AI-Model Ready

Objective: Reduce duplication of effort in making data AI-model ready, while providing the research community with tools needed to scale pre-processing work over time.

Approach: First, develop method to transform received formats into .txt and .mp3 files. Then, generate statistics on the datasets, to identify common themes, potential errors, biases, etc. Finally, clean the data by removing duplications, unnecessary formatting (e.g., xml notation, PII, data we do not have warrant to share, etc.) with as much automation as possible .

In Practice: Leveraging 5,000 scanned Chichewa newspaper pages and working with local students, build Chichewa OCR model for automating conversion of newsprint pdf to text. Working with the National Statistical Office, developing method to anonymize recorded household surveys.

```
# Step 5: Run Tesseract OCR
# You can adjust config for layout handling (e.g., --psm 1 or 3)
custom_config = r'--oem 3 --psm 3' # Use 3 for full page text
extracted_text = pytesseract.image_to_string(thresh, config=custom_config, lang='eng')

# Step 6: Display the Extracted Text
print("Extracted Text:\n")
print(extracted_text)

# Step 7: Save Text to File (Optional)
with open("ocr_output.txt", "w", encoding='utf-8') as f:
    f.write(extracted_text)

print("\n✓ OCR complete. Text saved to ocr_output.txt")
```

Original

Shasha
Janu-'worry'

Newspaper PDF Text Extraction

Thresholded (for OCR)

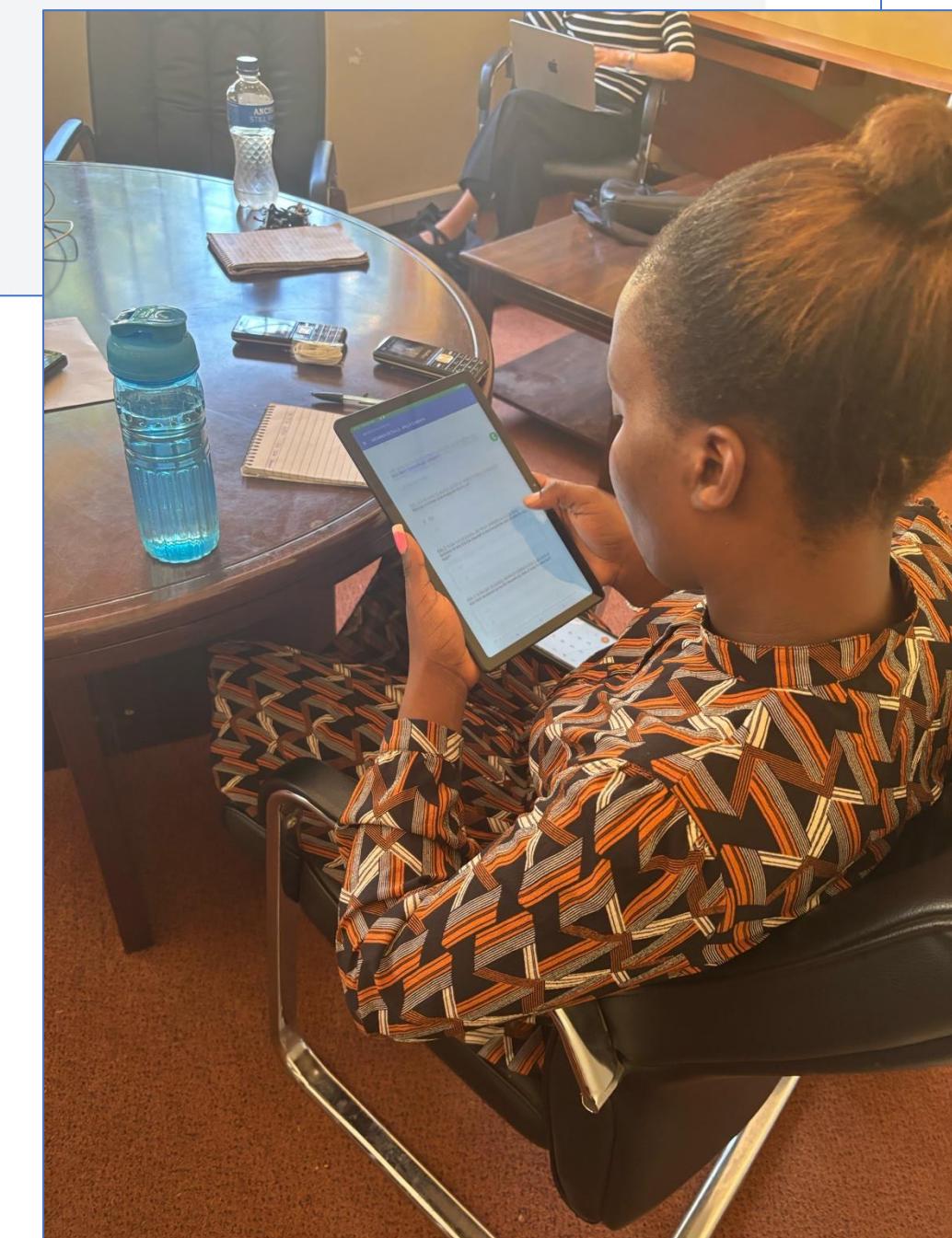
Shasha
Janu-'worry'

CHEBakali

Mkazi adandithera ndalamaza zanga

Extracted Text:

ZAKUKHOSI
anjekeya adatulutsa
N diso pa mtunda.
"Zimene mukuchitazi
sindingasekerere ngakhale
pang'ono... Pulopozo akuyenera
kudy deyile pano ndiponso..."





Pre-Processing

Protecting IP Rights

Objective: Mitigate risk of non-compliance with license agreements by limiting ability to download complete original copyright contents, without need to set up expensive and cumbersome self-contained trusted compute environments.

Approach: To prevent copyright infringement, chunk data based on section headers and file length (e.g., for text, page-section-paragraph-sentence groups) and shuffle. Add context summary to the meta data for each chunk, which will be callable via API. The API is being designed to include limiters and monitoring tools, to ensure original contents are not stitched together by users, as well as injection of 'canaries' that are assigned to each API-registrant, to support future license agreement enforcement.

Original Licensed Article

Drought Forces Thousands of Malawians to Abandon Their Homes

The Nation (Malawi) | February 8, 2024 | By Grace Mwale

Severe drought conditions in southern Malawi have displaced over 12,000 families in the past six months, forcing communities to abandon ancestral lands in search of water and livelihood opportunities.

Traditional leader Chief Ngabu from Nsanje district reported that three villages under his jurisdiction are now completely empty. "Our people have no choice but to move closer to the Shire River," he explained during a recent interview.

The displacement has created new challenges for host communities along the river valleys. Local schools report enrollment increases of 40%, while health clinics struggle with limited resources to serve the growing population.

Government response has been slow, with the Ministry of Disaster Management allocating only 2 billion kwacha for emergency assistance. Aid organizations estimate the actual need at 15 billion kwacha to provide adequate temporary shelter and food security.

Example: Processing Climate Displacement Article

Processed JSON Database Entry

```
{
  "article_id": "nation_mw_20240208_001",
  "source": "The Nation (Malawi)",
  "date": "2024-02-08",
  "author": "Grace Mwale",
  "language": "en",
  "region": "Southern Africa",
  "topics": ["climate_displacement", "drought", "migration"],
  "context_summary": "Drought displaces 12,000 families in Malawi",
  "chunks": [
    {
      "chunk_id": "displacement_impact",
      "content": "Severe drought conditions in southern Malawi have displaced over 12,000 families..."
    },
    {
      "chunk_id": "community_response",
      "content": "Traditional leader Chief Ngabu from Nsanje district reported..."
    },
    {
      "chunk_id": "host_community_strain",
      "content": "The displacement has created new challenges for host communities..."
    }
  ]
}
```

API Query Example

```
GET /api/content?
topics=climate_displacement & region=Southern_Africa & language=en & max_chunks_per_article=1
```

Response: Returns randomized chunks from multiple articles matching criteria. System ensures no single article can be reconstructed through multiple API calls.

Screenshot: API and IP Protection Concept (Claude.AI)





Library Management

Supporting Partnership Members and Data Fellows

Objective: Make access to the library seamless for partner international organization staff and Data Fellows

Approach: On the back end, leverage the Development Data Partnership's existing framework for receiving and reviewing proposals, securing agreement to data licensing terms, provisioning data access, providing supporting documentation and technical assistance, and monitoring usage.

The screenshot shows a project management interface for 'The Gambia sovereign flood insurance assessment'. At the top, there is a navigation bar with links for 'Data & Proposals', 'Proposal Submission', 'Documentation', and 'Blog'. A user profile for 'Holly Krambeck Admin' is also visible. The main content area displays the project title and a progress bar indicating the status of various milestones: 'Terms & Conditions' (0/2 Members Agreed, 09 Jun 2025), 'Admin Review' (Approved, 10 Jun 2025), 'Data Partner Review' (0/1 Approved), 'Data/Services Provisioning', 'Proposal Underway', and 'Completion Form'. Below the progress bar, the 'Proposal Details' section includes a summary of the project's purpose and a 'READ FULL PROPOSAL' button. The 'Team Member Agreement' section lists two team members: 'Felix Lung' (Manager, flung@worldbank.org) and 'Anne Hilger' (Task Team Lead, ahilger@worldbank.org), each with a 'VIEW TERMS & CONDITIONS' button. At the bottom, a footer bar shows 'Global Flood Hazard Maps', '0/2 MEMBER(S) AGREED', and 'ADMIN APPROVED'.

Screenshot: Development Data Partnership Project Management Portal



Library Management

Supporting Scaling

Objective: Define 1-3 solutions from different providers (AWS, Google Cloud, Azure) for scaling national libraries.

Approach: On the back end, prototype purpose-built data governance solutions for managing, fulfilling, and monitoring requests, leveraging support from technology partners. On the front end, create a public-facing application and meta-data catalog for researchers seeking use of the library.

The image displays three screenshots of the AWS Prototype Backend and Catalog Solution. The top-left screenshot shows the Data Catalog search interface with filters for 'Data type', 'Asset type', 'Owning project', and 'Domain unit'. The search results for 'Chichewa' are shown, with two assets listed: 'Chichewa Newprints Content' and 'Chichewa Radio Station Audio Recording'. The top-right screenshot shows the 'Subscription requests' page for the 'radio_mining-chichewa-project', listing incoming and outgoing requests. The bottom screenshot shows the 'Data' section of a project, displaying an S3 bucket structure for 'Chichewa Good News 17860/' and its contents, including subfolders like 'audio/' and 'chichewa/'. A detailed view of the 'Chichewa Good News 17860/' folder shows its 'Folder details' such as Name, S3 URI, ARN, Asset type (S3 Object Collection), Publish status (Not published), and Business name.

This screenshot provides a closer look at the 'Data' section of the AWS Prototype Backend and Catalog Solution. It shows the hierarchical structure of an S3 bucket named 'Chichewa Good News 17860/'. The bucket contains several subfolders: 'Lakehouse', 'AwsDataCatalog', 'Buckets', and 'S3 (project.s3_default_folder)'. Within 'S3 (project.s3_default_folder)', there are further subfolders like 'amazon-sagemaker-495599737147...', 'S3 (lang_content_chichewa.s3)', 'datalab-content-repo/', 'kenya/', 'malawi/', 'chichewa/', and 'audio/'. A specific item, 'Chichewa Good News 17860...', is selected, showing its details: Name (Chichewa Good News 17860/), S3 URI (s3://datalab-content-repo/malawi/chichewa/audio/Chichewa Good News 17860/), ARN (arn:aws:s3:::datalab-content-repo/malawi/chichewa/audio/Chichewa Good News 17860/), Asset type (S3 Object Collection), Publish status (Not published), and Business name (—).



Library Management

Guidance and Tools for Users

Objective: Mitigate risk of non-compliance with license agreements and reduce duplication of effort.

Approach: As part of the sub-licensing agreement, require researchers to submit results of a model test provided through the Partnership to ensure fine-tuned models minimize 'leakage' of original IP. In addition, provide researchers with other accompanying documentation in the Partnership's living repository of code examples and guidance.



Search x + K

Getting Started

- How it Works?
- Engage with the Community

Collections

- Climate
- Mobility
- Open Data
- Strategic Briefs

Documentation

- Data Partners
- Tutorials

Resources

- World Bank Data Lab
- Projects & News
- Project Template

Development Data Partnership Documentation

The [Development Data Partnership](#) brings together Companies and International Organizations to solve global development challenges.

 Development Data Partnership: Documentation Hub

Documentation Hub

Watch on YouTube



Community
Get involved, ask questions and share feedback, talk about ideas and new methodologies, and learn from the community

Documentation
Consult and contribute to documentation, tutorials, code snippets and examples focused in international development

Projects & News
Stay tuned and get inspired by projects and stories from Development and Data Partners

Contribute
We welcome you to join us in forming the Partnership and helping one another learn and build for public good

Screenshot: Development Data Partnership Documentation Portal

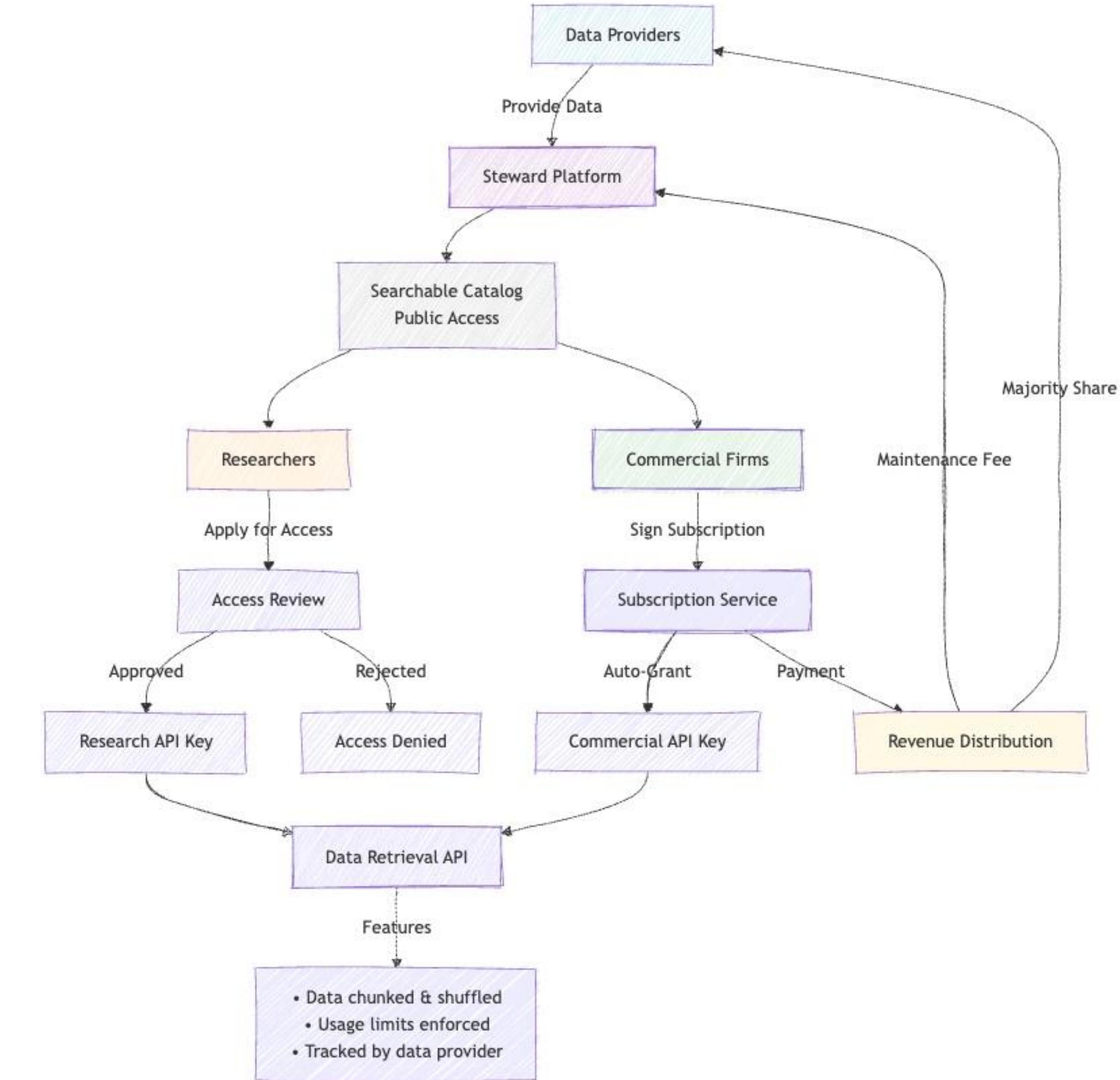


Scaling and Sustaining

Transferring Management to Country

Objective: Transfer ongoing stewardship to the national government for sustained content additions and revenue generation.

Approach: Learning by doing, through the Malawi pilot, work with Ministry of Information, University of Malawi, and MAREN (Malawi Research and Education Network) to establish a sustaining stewardship and funding model for library content provisioning, with different API tiers for research and commercial use, as well as a built-in sustained update and funding mechanism.





Scaling and Sustaining

Replicating Model with Other Governments

Objective: Build capacity of other governments to build their own libraries.

Approach: Prepare templates, guidance, and code accessible through a single open GitHub repository via living and dynamically-generated web book. Through the World Bank Digital Transformation investment team, the Partnership team has already received requests to support and fund replication of the Malawi program in Zambia and Uzbekistan, demonstrating the strong demand for these models and frameworks.



Search x + K

Curation

Library Curation Materials

Processing

Library Data Processing Materials

Processing Notebooks

Library Management and Governance

Library Governance and Management

Training Materials

Training Materials

Additional Resources

Development Data Partnership

Library Curation Materials

Following are materials to support acquisition and curation of proprietary content for building low-resource language digital libraries. The materials include playbooks and legal document templates for sourcing materials from government agencies and media houses, as well as tools and guidance for meta data schemas, catalog creation, and managing data requests.

Type	Material	Link
Content Acquisition	Language Content Workshop Planner, Invitation Templates, Stakeholder Briefs, and PPT Starter Deck	Workshop Guide
Legal	Content Negotiation Terms Sheet and Guide	Content Terms Sheet
Legal	Template Master Data License Agreement	Draft AI MDLA
Technical	Library Meta Data Schema and Template	Draft Schema
Technical	Library Meta Data Automation	
Technical	Catalog Set Up (DDP and Third Parties)	
Governance	Data Curation Guide	
Training	Complete Library Curation Course	

Audio Data Collection | Phone-Based Household Surveys

TYPE	MATERIAL	LINK
Technical	Survey Recording Manual	
Legal	Survey Recording Consent Agreement	HH Survey Consent Agreement

<https://bit.ly/data-for-AI-library>





Scaling and Sustaining

Replicating Model with Other Governments

Objective: Build capacity of other governments to build their own libraries.

Approach: Prepare templates, guidance, and code accessible through a single open GitHub repository via living and dynamically-generated web book. Through the World Bank Digital Transformation investment team, the Partnership team has already received requests to support and fund replication of the Malawi program in Zambia and Uzbekistan, demonstrating the strong demand for these models and frameworks.



The screenshot shows a website interface for 'Library Curation Materials'. At the top right are icons for refresh, download, and settings. Below the title 'Library Curation Materials' is a search bar with a magnifying glass icon and a 'K' button. The sidebar on the left lists categories: Curation (selected), Processing, LLM Guidance, Training Materials, and Additional Resources. Under 'Curation', there is a link to 'Library Curation Materials'. Under 'Processing', links include 'Library Data Processing Materials' and 'Processing Notebooks'. Under 'LLM Guidance', there is a link to 'LLM Guidance'. Under 'Training Materials', there is a link to 'Training Materials'. Under 'Additional Resources', there is a link to 'Development Data Partnership'. The main content area displays a table of materials categorized by Type (Technical, Management, Legal) and Material name.



Library Curation Materials

Following are materials to support acquisition and curation of proprietary content for building low-resource language digital libraries. The materials include playbooks and legal document templates for sourcing materials from National Statistical Offices and media houses, as well as tools and guidance for meta data schemas, catalog creation, and managing data requests.

Type	Material	Link
Technical	Pipeline Design	
Management	Media House Workshop Planner	
Legal	Content Negotiation Terms Sheet	
Legal	Template Master Data License Agreement	
Technical	Survey Recording Manual	
Legal	Survey Recording Consent Agreement	
Technical	Survey Transcription Manual	
Technical	Library Meta Data Schema	
Technical	Library Meta Data Automation Scripts	
Management	Usage Logging and Monitoring System	
Technical	Catalog Set Up	
Content	Links to Open Source Content for Integration	
Management	Catalog Management	

<https://bit.ly/data-for-AI-library>



Component 2

Unlocking Grantee Datasets





Unlocking Grantee Data

Component 2 | Grantee Data

Component Objective

To create a scalable, replicable model for provisioning access to non-open grantee datasets for public good.

Challenge

The Gates Foundation funds dozens of organizations each year to support data collection, and while dissemination is a requirement, grantees often struggle with the responsibility of data provisioning months and years after the grant has closed. With no unified catalog, legal agreement, or data architecture, third parties are challenged to discover or access these potentially useful datasets.

Solution

Leverage Partnership legal, governance, and technical frameworks to unlock data from Gates Foundation grantees for use in public good research and AI applications.



Approach: Learning by Doing

Component 2 | Grantee Data

1

Legal Framework | Curation

The team will engage and source content from Gates Foundation grantees, leveraging a modified version of the DDP master data license agreement. The team will leverage the existing DDP infrastructure to ingest, store, and provision datasets.

2

Technical Framework | Pre-Processing

As is needed, the team may engage in pre-processing of datasets to support ease of use, as well as preparation of APIs and/or documentation.

3

Technical Framework | Library Management

The team will adjust the DDP data application and provisioning process for the project. The team will prepare third-party engagement and data provisioning workflow that is more easily scalable than the current model.

4

Governance Framework | Scaling & Sustaining

Oversight will be managed through the Partnership's current inclusive governance framework. The team will support an outreach effort across the international organization partners for awareness building and training on using the hosted datasets and may also support a kick-off hack-a-thon style event to encourage more third-party researchers to apply for and use the datasets.



Project Timeline

Key Milestones





Project Management

How We Organize Our Work

The core team includes the Development Data Partnership team, World Bank legal counsel and data scientists, and consultants with specialized experience in working in the Malawi context and with low-resource languages. The extended team include colleagues from the World Bank IT department, as well as the Development Economics Data Group and Digital Transformation teams.

Project tasks are organized and monitored using a GitHub task tracker, visible to team members for transparent project management.

Project GitHub Task Tracker

GF-data-library						
Backlog	Team capacity	Roadmap	My items	Current iteration	+ New view	Add
Assignees						
claudiacalderon	2					
dmatekenya	10					
farhanreynaldo	2					
fenimi	8					
Holly-Transport	20					
SahitiSarva	2					
No Assignees	19					
Filter by keyword or by field						
Title	...	Status	...	Labels	...	Assignees
1 Curation Pipeline Design #2	In Progress	In Progress	In Progress	O-Low-Resource-Language		dmatekenya and f...
2 Curation Pipeline Design #2	In Progress	In Progress	In Progress	O-Grantee and Research Data		farhanreynaldo an...
3 Curation GF Grantee Dataset Identification #14	In Progress	In Progress	In Progress	O-Grantee and Research Data		Holly-Transport
4 Curation SUB TASK RDA Technical and Legal Backgrounder #15	In Progress	In Progress	In Progress	O-Grantee and Research Data		Holly-Transport
5 Curation Media House Workshop Planner #33	In Progress	In Progress	In Progress	O-Low-Resource-Language		Holly-Transport
6 Curation Content Negotiation Terms Sheet #32	In Progress	In Progress	In Progress	O-Low-Resource-Language		Holly-Transport
7 Curation Prepare Standard Master Data License Agreement (MDLA) #3	In Progress	In Progress	In Progress	O-Grantee and Research Data		Holly-Transport
8 Curation Prepare Standard Master Data License Agreement (MDLA) #3	In Progress	In Progress	In Progress	O-Grantee and Research Data		Holly-Transport
9 Curation Transcription Manual #4	In Progress	In Progress	In Progress	O-Low-Resource-Language		dmatekenya
10 Curation Survey Recording Manual #5	In Progress	In Progress	In Progress	O-Low-Resource-Language		dmatekenya
11 Curation Links to Open Source Low-Resource Language Datasets #6	In Progress	In Progress	In Progress	O-Low-Resource-Language		fenimi
12 Curation MetaData Schema #7	Todo	Todo	Todo	O-Low-Resource-Language		
13 Curation MetaData Schema #4	Todo	Todo	Todo	O-Grantee and Research Data		
14 Curation MetaData Auto Tagging Scripts #8	Todo	Todo	Todo	O-Low-Resource-Language		
15 Curation MetaData Auto Tagging Scripts #5	Todo	Todo	Todo	O-Grantee and Research Data		
16 Curation Monitoring Checking Out and Returns - Initial Design #26	In Progress	In Progress	In Progress	O-Low-Resource-Language		Holly-Transport
17 Curation Monitoring Checking Out and Returns - Initial Design #9	Todo	Todo	Todo	O-Grantee and Research Data		dmatekenya
18 Curation Chunking, Shuffling, and API Design #43	Todo	Todo	Todo	O-Low-Resource-Language		
19 Curation Integrate Library with Data Partnership Catalog and Documentation #9	Todo	Todo	Todo	O-Low-Resource-Language		Holly-Transport an...
20 Curation Integrate Library with Data Partnership Catalog and Documentation #6	Todo	Todo	Todo	O-Grantee and Research Data		Holly-Transport an...
21 Curation Design External Meta Data Catalog #10	Todo	Todo	Todo	O-Low-Resource-Language		
22 Curation Design External Meta Data Catalog #7	Todo	Todo	Todo	O-Grantee and Research Data		

The Core Team

Holly Krambeck
Dunstan Matekenya
Yoon Jee Kim
Keongmin Yoon
Aivin Solatorio
Sahiti Sarva
Maria Sol Tadeo
Claudia Calderon
Evance Mathewe
Ifeoluwanimi Esther Ogbeba
Towera Moyo

Task Team Leader
Data Scientist
Legal Counsel
Legal Counsel
Data Scientist - Advisor
Data Scientist - DDP
Data Scientist - DDP
Partnership Specialist - DDP
Consultant – Malawi Pilot
Consultant – Data Science
Consultant – Data Science



Milestones

February 2025 – June 2026

	Q1-3 2025	Q4 2025	Q1 2026	Q2 2026	Q3 2026
Curation	<ul style="list-style-type: none">• Data owner consultations• Legal terms sheet	<ul style="list-style-type: none">• Legal agreements• Data collection• Metadata schema	<ul style="list-style-type: none">• Curation Documentation Completion	<ul style="list-style-type: none">• Continued data collection / curation	<ul style="list-style-type: none">• Continued data collection / curation
Processing		<ul style="list-style-type: none">• Processing documentation• Processing implementation	<ul style="list-style-type: none">• Data privacy and warranty checks	<ul style="list-style-type: none">• Processing Documentation Completion	<ul style="list-style-type: none">• Continued data processing
Management	<ul style="list-style-type: none">• LRL government endorsement	<ul style="list-style-type: none">• Grantee dataset application procedure	<ul style="list-style-type: none">• Grantee library launch• API Solution Developed	<ul style="list-style-type: none">• LRL model 'leakage' verification tool• LRL LLM/ASR guidance• LRL Library launch	<ul style="list-style-type: none">• Library Management Documentation Completion
Capacity Building & Scaling					<ul style="list-style-type: none">• Training Course• Complete Toolkit



THANK YOU

Development Data Partnership