METADATA SCHEMAS FOR LOW RESOURCE LANGUAGE LIBRARY

Introduction

The Development Data Partnership and the Gates Foundation aim to create a scalable and replicable model for building national digital language libraries to support the development of AI solutions. This involves curating datasets in low-resource languages. A pilot national library is being developed in Malawi to host datasets in Chichewa, a language spoken by over 20 million people. These datasets will be drawn from news, books, radio, and survey sources, all collected under clear consent, strong privacy protections, and non-commercial terms. The pilot is expected to result in the curation of hundreds of thousands of news articles, as well as tens of thousands of hours of audio and video.

The project seeks to ensure that the datasets produced within these libraries are as discoverable and usable as possible for AI model training and fine-tuning. One of the keyways this is being achieved is through the collection of complete metadata. Metadata provides structured descriptions that explain what a dataset contains, how it was produced, its format, provenance, and the conditions under which it can be used, all of which are recognized as essential for enabling data reuse and interoperability (Greenberg, 2005; Wilkinson et al., 2016).

Comprehensive metadata is particularly important in this project because it focuses on low-resource languages, where every dataset carries significant value. Well-documented Chichewa data can support a wide range of research activities, but only if it is easy to find and includes sufficient technical detail to allow others to use it confidently, without needing to contact the creators for basic clarification. In this context, metadata serves as a critical link between dataset creators and potential users for AI model training and fine-

tuning, supporting transparency, trust, and reproducibility in AI systems (Gebru et al., 2021). For example, audio datasets require metadata describing sampling rate, signal-to-noise ratio, duration, and recording conditions for researchers to assess suitability for specific training tasks.

A common approach to documenting metadata is through the use of established metadata standards. Metadata standards provide shared rules and controlled vocabularies that enable consistent interpretation of data across systems and communities, improving discoverability, interoperability, and long-term reuse (Weibel et al., 1998; ISO, 2017). No single existing standard, however, fully met the project's requirements. A review of general standards such as Dublin Core, domain-specific standards such as EBUCore for broadcast media, and machine learning–specific standards such as Croissant found that multiple standards were required. Consequently, a combined approach was adopted, supported by a custom namespace to capture development-specific attributes while maintaining compatibility with existing vocabularies (Heery & Patel, 2000).

This document outlines the objectives of the metadata, describes the approach used to integrate multiple metadata standards, and presents the resulting schemas. These schemas are intended to guide the collection of our metadata and to serve as a useful resource for other low-resource language dataset initiatives.

Main Objectives

The metadata collected was guided by four core objectives: ensuring that the datasets are discoverable, usable, reproducible, and aligned with ethical practices.

Discoverability

Our objective was to ensure that datasets can be found and correctly identified by users or systems through their metadata. Good discoverability means that resources can be easily located through relevant searches because the metadata uses commonly searched terms,

aligns with shared vocabularies, and captures meaningful descriptive information (Weibel et al., 1998; Wilkinson et al., 2016).

This objective was achieved by labelling datasets with key attributes such as title, language, modality, and domain, enabling search tools to index and return them accurately (Greenberg, 2005).

Usability

Our aim was to provide sufficient information for users to effectively work with the datasets once they have been discovered. This includes understanding what the dataset contains, how it can be used, and whether it is suitable for developing AI tools. Good usability allows users to quickly assess whether a resource meets their needs because key details are clearly documented upfront (Bruce & Hillmann, 2004; Gebru et al., 2021).

This was achieved by documenting dataset characteristics relevant to machine learning use, including information on data splits, preprocessing steps, and annotation practices, as well as technical characteristics such as word count, token count for text datasets. Poor usability occurs when users can locate a resource but cannot determine whether it is appropriate for their purposes, often resulting in wasted time downloading datasets that ultimately prove unsuitable (Gebru et al., 2021).

Reproducibility

Another key objective of the metadata was to document sufficient detail to allow others to replicate, verify, or build upon the work, creating a clear audit trail from raw data to final outputs. Good reproducibility metadata enables others to understand and repeat how a dataset was created by clearly documenting methods, provenance, and processing steps (Hedstrom & Lee, 2002; Pineau et al., 2021).

This was achieved through transparent descriptions of data collection and processing workflows. With this information, researchers can follow and understand the steps taken to

transform hundreds of thousands of raw audio files into a cleaned, deduplicated, and quality-assessed dataset suitable for AI training.

Ethical Practices

Our goal was also to transparently document information necessary for responsible use and compliance with ethical obligations. Good ethical practice includes clear documentation of data provenance, how the data was collected, by whom, and under what conditions, enabling researchers to assess potential biases and determine whether a dataset is appropriate for their intended use (Gebru et al., 2021).

This objective was met by collecting information related to licensing to support legal compliance, as well as details on consent and privacy measures to ensure that ethical standards are upheld.

Identifying a Metadata Standard

The standards-based approach to metadata documentation was guided by the following main considerations:

- **Shared semantics:** Standards provide consistent meanings for metadata elements across systems and disciplines, reducing semantic ambiguity and preventing misinterpretation when data are exchanged (Greenberg, 2005).
- **Interoperability:** Standardized metadata enables automated discovery, integration, and processing across repositories and platforms, whereas non-standard metadata requires manual mapping and increases cost and error (Weibel et al., 1998; ISO, 2017).
- **Long-term preservation:** Metadata created using recognized standards remains intelligible over time as technologies, staff, and institutions change, unlike locally documented metadata that depends on tacit knowledge (Hedstrom & Lee, 2002).

- **Quality and consistency:** Standards define required elements, structures, and controlled vocabularies, reducing omissions and inconsistencies common in free-text documentation and supporting reliable reuse (Bruce & Hillmann, 2004).

Applying these considerations, the team reviewed existing metadata standards and found that no single schema captured all the information required to meet the objectives of the project discussed above. No individual standard provided sufficient coverage of language identification, rights and licensing, data governance, provenance, and the technical properties needed for AI training. Some standards offered strong support for linguistic description but lacked AI-relevant technical fields, while others focused on machine learning workflows but provided limited detail on consent, privacy, or data collection context. To address these gaps, the team adopted an application profile approach.

Application Profile Methodology

The application profile approach, proposed by Heery and Patel (2000), provides the theoretical foundation for the schema design used in this project. Application profiles combine data elements from one or more namespaces (standards), selected and adapted by implementers to meet specific local application needs (Coyle & Baker, 2009). This approach enables the integration of fields from multiple standards into unified, interoperable schemas that address project requirements across text, audio, and video datasets.

Application profiles allow selected elements to be constrained or extended to suit local needs, ensure that design decisions and mappings are clearly documented, and support implementations that reflect real use cases while balancing completeness with practical effort (Heery & Patel, 2000; Nilsson et al., 2008). Where existing standards do not fully address project requirements, application profile methodology supports the creation of a custom namespace (Coyle & Baker, 2009).

Following this approach, an application profile was developed for the project by balancing standardization with domain-specific and technical requirements. The profile draws on

nine established standards. Rather than adopting these standards in their entirety, only relevant properties were selectively adapted, with a focus on elements critical to the objectives discussed above and to low-resource language contexts. Where gaps remained, the framework was extended through a custom public namespace designed to address project-specific requirements without compromising compatibility with established vocabularies, consistent with established application profile practice (Nilsson et al., 2008). This namespace was documented on a dedicated web page, including definitions, explanations, and example usage for each custom term. The namespace will be maintained by the Development Data Partnership team at the World Bank.

The resulting application profile and the standards it incorporates are described in the following section.

Low-Resource Language Metadata Application Profile

The metadata collected in this application profile falls into five main categories: descriptive, technical, structural, administrative, and operational metadata. The profile primarily extends the Croissant standard, developed by MLCommons to improve the discoverability, interoperability, and reuse of machine learning datasets, because it provides native support for describing dataset formats, contents, and machine learning–specific elements such as splits, features, and labels (MLCommons, 2024). Croissant also provides properties for documenting responsible AI considerations, such as intended use, and known limitations. Importantly, Croissant builds on the Schema.org Dataset vocabulary, a widely adopted framework that provides machine-readable properties for describing structured data on the web and is currently used to document millions of datasets (Guha et al., 2016).

The application profile incorporates Schema.org properties beyond those specified in the Croissant standard, as well as elements from multiple complementary standards:

- **Dublin Core Terms** - A widely adopted metadata standard maintained by the Dublin Core Metadata Initiative. It provides 15 core elements such as creator, title, subject,

and date, offering a simple and widely recognized way to describe resources across digital libraries, archives, and repositories (Weibel et al., 1998; ISO, 2017).

- **Data Catalog Vocabulary (DCAT) -** A W3C standard designed for describing datasets in data catalogues. It supports interoperability between different data portals, allowing datasets to be harvested and incorporated into broader data discovery platforms (W3C, 2020).
- **Data Quality Vocabulary (DQV)** - A W3C standard used to express data quality metrics and assessments. It captures aspects such as accuracy, completeness, and consistency, which are necessary for AI practitioners when judging whether a dataset is suitable for training (W3C, 2016).
- **Provenance Ontology (PROV-O)** - A W3C standard for recording the origin and history of data throughout its lifecycle. It documents who created or modified the data, when and how it was handled, and which sources it drew from, all of which are central to reproducibility and trust in AI datasets (W3C, 2013).
- **EBUCore** - The European Broadcasting Union's metadata standard for audio and video. It includes detailed technical information such as codecs, bit rates, and frame rates, along with content descriptions covering languages, subtitles, and accessibility features (EBU, 2018).
- **Open Language Archives Community (OLAC)**- A metadata standard tailored to language resources and linguistic data archives. It includes language-specific descriptive elements (ISO 639 language codes, linguistic data types, discourse types) that general standards do not cover (Bird & Simons, 2003).
- **Bibliographic Ontology (BIBO)** – A metadata standard for describing bibliographic resources, including books, articles, reports, and other scholarly materials. It provides structured properties for citations, publication details, authorship, and document types, supporting consistent documentation and linking of academic and research outputs (D'Arcus & Giasson, 2009).

- **Data Documentation Initiative (DDI)** - An international standard for describing structured data collections. It provides detailed elements for variables, coding schemes, methodology, and collection protocols commonly used in social science and survey research (Vardigan et al., 2008).
- **RDF Schema (RDFS)** - The core vocabulary for defining relationships and hierarchies in RDF data. It provides basic elements such as labels, comments, and class definitions that make metadata machine-readable and semantically linked (W3C, 2014).

The custom namespace, Development Data Partnership Vocabulary (DDPV) captures information relevant to the project such as data governance terms, participant characteristics including age range and dialect information, and technical text metadata relevant to the development of AI solutions.

Using this application profile, we developed metadata schemas for three primary data types: audio, text, and video. The following section presents the schemas, documenting the terms selected, their source standards, definitions, cardinality requirements, and collection methods.

Schemas

This section first presents the general schema outline, which consists of metadata attributes common to all schemas, and then the specific metadata elements for each data type.

The table below lists the prefixes assigned to each metadata standard included in the schemas.

Standard Name	**Prefix**
Croissant	cr

Croissant Responsible AI Terms	rai
Schema.org	sc
Dublin Core Terms	dct
Data Quality Vocabulary	dqv
Data Documentation Initiative	ddi
Provenance Ontology	prov
Open Language Archives Community	olac
Bibliographic Ontology	bibo
RDF Schema	rdfs
Development Data Partnership Vocabulary	ddpv

General Schema

Field Name	Standard	Purpose	Expected Type	Cardinality	Entry Method	Requirement
Dataset Identity						

@context	JSON-LD	Declares the standards used in the schema (Croissant, Schema.org, DQV, PROV-O, DDI, EBUCore). It helps tools know which vocabulary each property belongs to	Array of URL and/or Object	MANY	Manual	Mandatory
@type	JSON-LD	Declares the type of object being described (e.g Dataset)	Text	ONE	Manual	Mandatory
sc:identifier	Schema.org	A unique identifier for the dataset	Text	ONE	Automatic	Mandatory
ddi:altID	DDI	This repeatable element is used to enter identifiers other than the primary ID. It can for example, be a	Text	MANY	Manual	Optional

		Digital Object Identifier (DOI)				
dct:conformsTo	Dublin Core Terms	Declares the standard to which the dataset conforms	URL	ONE	Manual	Mandatory
Basic Dataset Metadata						
sc:name	Schema.org	The dataset name	Text	ONE	Manual	Mandatory
dct:alternative	Dublin Core Terms	An alternate dataset name, possibly an abbreviated version	Text	ONE	Manual	Optional
sc:description	Schema.org	Summary of dataset contents and purpose	Text	ONE	Manual	Mandatory
sc:url	Schema.org	Landing page for the dataset	URL	ONE	Manual	Mandatory
sc:version	Schema.org	Dataset version identifier	Text	ONE	Manual	Mandatory

sc:datePublished	Schema.org	The date the dataset was published	Date / Date and Time	ONE	Manual	Mandatory
sc:keywords	Schema.org	A set of keywords associated with the dataset	Text	MANY	Manual	Mandatory
sc:dateCreated	Schema.org	The date the dataset was initially created	Date / Date and Time	ONE	Manual	Mandatory
sc:dateModified	Schema.org	Last modified timestamp for the dataset	Date / Date and Time	ONE	Manual	Mandatory
cr:isLiveDataset	Croissant	Whether the dataset is a live dataset	Boolean	ONE	Manual	Optional
dcat:theme	DCAT	High-level categories or subjects describing the dataset (controlled vocabulary URIs preferred)	URLs or Text	MANY	Manual	Recommended

dcat:themeTaxonomy	DCAT	URI of the controlled vocabulary or taxonomy defining the dataset themes	URL	ONE	Manual	Optional
Language & Identifiers						
olac:subjectLanguage	OLAC	Language(s) represented or studied in the dataset, using ISO 639-3 or BCP-47 codes	Object with {name, code}	MANY	Manual	Mandatory
sc:inLanguage	Schema.org	Language(s) used to describe or present the dataset, following BCP-47 tags	Object with {name, identifier}	MANY	Manual	Mandatory
olac:linguisticType	OLAC	Type of linguistic resource	Controlled list of values (Enum)	ONE	Manual	Optional

olac:discourseType	OLAC	Genre or discourse style of the content	Text	MANY	Manual	Optional
Geographic & Time Coverage						
sc:temporalCoverage	Schema.org	Time range covered by the data	Text	ONE	Manual	Mandatory
sc:spatialCoverage	Schema.org	Geographic scope	Text	ONE	Manual	Mandatory
People and Organizations						
sc:creator	Schema.org	Organization/team that created the dataset	Organization/ Person	ONE	Manual	Mandatory
sc:provider	Schema.org	Organization providing/hosting the dataset	Organization/ Person	ONE	Manual	Mandatory
sc:sourceOrganization	Schema.org	The entities from which the dataset was obtained or derived	Organization / Person	ONE	Manual	Mandatory

sc:contributor	Schema.org	Details of person or organization that contributed to the creation of the dataset	Organization / Person	MANY	Manual	Optional
sc:funder	Schema.org	Organization or person providing financial support	URL or Object	MANY	Manual	Optional
sc:audience	Schema.org	Intended audience for an item, such as the group for whom the dataset is designed	Text	ONE	Manual	Optional
sc:contactPoint	Schema.org	Provides the contact details of individuals or organizations responsible for dataset inquiries, access requests, or further information	Contact Point	MANY	Manual	Recommended
Rights & Licensing						

sc:license	Schema.org	Legal license under which the dataset is shared	Text / URL	ONE	Manual	Mandatory
sc:copyrightHolder	Schema.org	The party holding the legal copyright to the dataset	Text	ONE	Manual	Optional
sc:copyrightNotice	Schema.org	Describing the copyright aspects of the dataset	Text	ONE	Manual	Optional
sc:copyrightYear	Schema.org	The year during which the claimed copyright for the dataset was first asserted	Number / Text	ONE	Manual	Optional
sc:usageInfo	Schema.org	Instructions or context on how the dataset can or should be used beyond the formal license	CreativeWork / URL / Text	MANY	Manual	Optional

sc:conditionsOfAccess	Schema.org	Dataset restrictions	Text	MANY	Manual	Optional
Privacy and Sensitivity						
ddpv:piiScreening	DDPV	States whether the dataset contains personally identifiable information	Boolean	ONE	Manual	Mandatory
ddpv:piiScreeningMethod	DDPV	Method used to remove PII (If yes to PII screening)	Enum	ONE	Manual	Recommended
ddpv:piiNotes	DDPV	Short note describing the type of PII present (if any)	Text	ONE	Manual	Recommended
ddpv:attribution	DDPV	Whether users must attribute	Boolean	ONE	Manual	Optional
ddpv:thirdPartyRestrictions	DDPV	Any publisher/broadcaster limitations	Text	MANY	Manual	Optional

ddpv:sensitiveContent	DDPV	Whether the dataset contains restricted or potentially harmful content	Boolean	ONE	Manual	Mandatory
ddpv:sensitiveNotes	DDPV	The types of sensitive content the dataset contains	Text	ONE	Manual	Recommended
ddpv:retentionPolicy	DDPV	Retention/review policy	Text	MANY	Manual	Optional
Responsible AI properties						
rai:dataCollection	Croissant RAI	Description of how the data was collected, including method, setting, and sources	Text or URL	ONE	Manual	Recommended
rai:dataCollectionType	Croissant RAI	Type or method of data collection	Enum or Array	MANY	Manual	Optional
rai:dataCollectionRawData	Croissant RAI	Source of the raw data before any processing	Text or URL	ONE	Manual	Recommended

rai:dataCollectionTimeFrameStart	Croissant RAI	Start date/time when data collection began	Date or DateTime	ONE	Manual	Optional
rai:dataCollectionTimeFrameEnd	Croissant RAI	End date/time when data collection was completed	Date or DateTime	ONE	Manual	Optional
rai:dataUseCases	Croissant RAI	Intended use cases for the dataset	Text or Array	MANY	Manual	Optional
rai:dataBiases	Croissant RAI	Known or potential biases in the dataset	Text or Array	MANY	Manual	Optional
rai:dataLimitations	Croissant RAI	Known limitations, gaps, risks, or caveats	Text or Array	MANY	Manual	Recommended
rai:dataSocialImpact	Croissant RAI	Description of the anticipated or potential social impact	Text	ONE	Manual	Optional
rai:dataReleaseMaintenancePlan	Croissant RAI	How the dataset will be maintained,	Text	ONE	Manual	Optional

		updated, and released				
rai:dataAnnotationProtocol	Croissant RAI	Protocol or guidelines used for data annotation/labeling	Text or URL	ONE	Manual	Optional
rai:dataAnnotationPlatform	Croissant RAI	Platform or tool used for annotation	Text	ONE	Manual	Optional
rai:dataAnnotationAnalysis	Croissant RAI	Analysis of annotation quality, inter-rater reliability	Text	ONE	Manual	Optional
rai:annotationsPerItem	Croissant RAI	Number of annotations collected per data item	Number	ONE	Manual	Optional
rai:dataPreprocessingProtocol	Croissant RAI	Steps taken to preprocess, clean, or filter the data	Text, URL, or Array	MANY	Manual	Optional

rai:dataManipulationProtocol	Croissant RAI	Description of how data was transformed, augmented, or manipulated	Text or Array	MANY	Manual	Optional
Data Quality						
dqv:hasQualityMeasurement	DQV	Aggregate technical/quality metrics for the dataset.	Array of Objects	MANY	Automatic	Optional
Provenance						
prov:qualifiedGeneration	PROV-O	The process or pipeline that created the dataset	Object Definition	ONE	Manual/Automatic	Recommended
prov:used	PROV-O	Inputs used in creation	URL	MANY	Manual/Automatic	Recommended
prov:wasDerivedFrom	PROV-O	Links a dataset to another dataset or source from which it was derived or transformed	Text / URL / Object	MANY	Manual/Automatic	Recommended

prov:wasAttributedTo	PROV-O	Links a dataset directly to the person, team, or organization responsible for it	URL	MANY	Manual/Automatic	Optional
prov:wasAssociatedWith	PROV-O	Teams or systems that were involved in the activities that produced the dataset	URL / Object	MANY	Manual/Automatic	Optional
Distribution and File Structure Distributions describe how a dataset's data can be accessed. In Croissant, this is expressed using FileObject for individual files (each with its own metadata) and FileSet for collections of related files (e.g., a directory of text files grouped together).						
distribution	Croissant	Canonical list of FileObjects or FileSets that make up the dataset	Array	MANY	Manual	Mandatory
cr:recordSet	Croissant	Defines the logical table of data records	Array of RecordSet	MANY	Manual	Mandatory

		within the dataset				
cr:fileObject	Croissant	List of individual files in the dataset	Array of FileObject	MANY	Manual	Optional
cr:fileSet	Croissant	Groups of related files that form one logical resource	Array of FileSet	MANY	Manual	Optional
File Object: Core Basic Properties						
@id	Croissant	Local identifier for cross-references	Text	ONE	Manual	Mandatory
sc:name	Schema.org	The name of the file	Text	ONE	Manual / Automatic	Mandatory
sc:contentUrl	Schema.org	Direct download URL for this file	URL	ONE	Manual	Mandatory
sc:encodingFormat	Schema.org	Identifies the overall file/container	Text	ONE	Manual / Automatic	Mandatory

		format using a MIME type				
cr:sha256	Croissant	SHA-256 checksum of the file content, used to verify file integrity and detect corruption	Text	ONE	Automatic	Recommended
sc:contentSize	Schema.org	File size in bytes	Number / Text	ONE	Automatic	Mandatory
sc:sameAs	Schema.org	URL (or local name) of a FileObject with the same content, but in a different format	URL	MANY	Manual	Optional
sc:dateModified	Schema.org	Last modified timestamp for the file	Date / Date and Time	ONE	Automatic	Mandatory
sc:inLanguage	Schema.org	Primary language of each file	Text	ONE	Automatic	Mandatory
dcat:theme	DCAT	High-level categories/sector	Text / URI	MANY	Manual	Mandatory

		s describing the file				
dcat:themeTaxonomy	DCAT	URI of controlled vocabulary defining themes	URI	ONE	Automatic	Optional
sc:keywords	Schema.org	Keywords or tags for the file	Text	MANY	Manual / Automatic	Mandatory
File Set						
cr:containedIn	Croissant	Points to the FileObject that physically contains the files (e.g., a TAR/ZIP)	Object	MANY	Manual	Optional
cr:includes	Croissant	Specifies which files to include	Object	MANY	Manual	Mandatory
cr:excludes	Croissant	Specifies which files to exclude	Object	MANY	Manual	Optional
sc:encodingFormat	Schema.org	Identifies the overall file/container format using a MIME type	Text	ONE	Manual	Mandatory

sc:dateModified	Schema.org	Last modified timestamp for the set's definition	Date / Date and Time	ONE	Automatic	Mandatory
Record Set Properties						
cr:key	Croissant	Unique key identifying each record within the RecordSet	Object	ONE	Manual	Mandatory
cr:field	Croissant	Defines the structure of the record by listing all its fields	Array of field objects	MANY	Manual	Mandatory
cr:records	Croissant	Contains or links to the actual recorded entries in the dataset	Array of record objects	MANY	Manual / Automatic	Mandatory
Field A Field is an element of a RecordSet. It can represent a table column, a nested attribute, or even another RecordSet in hierarchical data.						
cr:name	Croissant	Human-readable name of the field	Text	ONE	Manual	Mandatory

cr:dataType	Croissant	Defines the datatype of the field	Enum	ONE	Manual	Mandatory
cr:source	Croissant	Specifies the source file, column, or data element from which this field is derived	Text / URL	ONE	Manual	Mandatory
cr:equivalentProperty	Croissant	Maps the field to a standard property from another vocabulary	URL / Object	MANY	Manual	Optional
ML Attributes Under the record set, we can define ML-specific attributes.						
cr:split	Croissant	Indicates which records belong to the training, validation, or test set	Enum	MANY	Automatic	Optional

cr:label	Croissant	Declares the label or target variable for supervised learning	Text	MANY	Manual / Automatic	Optional
cr:content	Croissant + Schema.org	Identifies the input data (e.g., text, image, or audio) used by the model	Text / URL	ONE	Manual / Automatic	Optional
Citation and References						
cr:citeAs	Croissant	A citation to the dataset itself, or a citation for a publication that describes the dataset	Text / URL	ONE	Manual	Recommended
sc:isReferencedBy	Schema.org	External works (web pages, papers) that cite or reference this dataset	URL or Array of URLs	MANY	Manual	Optional
Additional Metadata						

sc:measurementTechnique	Schema.org	Measurement or sampling technique used	Text or URL	ONE	Manual	Optional
sc:interactionStatistic	Schema.org	User interaction counters (downloads, views, likes)	Array of Objects	MANY	Automatic	Optional
metadata_information	Metadata Editor Fields	Metadata about the Metadata record itself	Object	ONE	Manual	Mandatory

Text Schema Attributes

Field Name	**Standard**	**Purpose**	**Expected Type**	**Cardinality**	**Entry Method**	**Requirement**
dct:title	Dublin Core Terms	Title of the article associated with the file	Text	ONE	Manual	Mandatory
sc:author	Schema.org	Author(s) of the article or creative work	Text / Person / Organization	MANY	Manual	Recommended
dct:subject	Dublin Core Terms	Topic, domain, or genre of text content	Text	MANY	Automatic	Mandatory
dct:type	Dublin Core Terms	Type of document or resource	Text / URI	ONE	Automatic	Optional

sc:datePublished	Schema.org	Date when article was first published	Date / Date Time	ONE	Automatic	Mandatory
sc:wordcount	Schema.org	Total number of words in the text	Integer	ONE	Automatic	Optional
ddpv:charCount	DDPV	Total number of characters	Integer	ONE	Automatic	Optional
ddpv:tokenCount	DDPV	Total number of tokens	Integer	ONE	Automatic	Recommended
ddpv:tokenizerName	DDPV	Name of the tokenizer used to compute token counts	Text	ONE	Automatic	Recommended
ddpv:perplexityScore	DDPV	Model prediction quality score (lower = better)	Number	ONE	Automatic	Optional

Audio Schema Attributes

Field Name	**Standard**	**Purpose**	**Expected Type**	**Cardinality**	**Entry Method**	**Requirement**
ebucore:hasTranscrip	EBUCore	Transcript or subtitle file for the audio	Text	ONE	Manual/ Automatic	Recommended
ebucore:signalToNoiseRatio	EBUCore	Ratio of signal strength to background noise	Number	ONE	Automatic	Recommended

ebucore:duration	EBUCore	Length of individual media files	Number	ONE	Automatic	Recommended
ddpv:totalDurationHours	DDPV	Total hours of content	Number	ONE	Automatic	Recommended
ebucore:loudnessLUFS	EBUCore	Perceived loudness level measured by Integrated LUFS	Number	ONE	Automatic	Recommended
ebucore:sampleRate	EBUCore	Number of times per second the audio waveform is captured.	Number	ONE	Automatic	Optional
ebucore:bitDepth	EBUCore	How precisely each sample's amplitude is stored.	Number	ONE	Automatic	Optional
ebucore:bitrate	EBUCore	Overall data rate of the encoded audio. Indicates storage/bandwidth cost and perceived quality	Number	ONE	Automatic	Optional
ebucore:channels	EBUCore	How many independent audio channels are in each recording	Number	ONE	Automatic	Optional

ebucore:audioCodec	EBUCore	Audio compression format (e.g., AAC, MP3, PCM, AC-3)	Text	ONE	Automatic	Recommended
ddpv:equipmentType	DDPV	Recording device or microphone model	Text	ONE	Automatic	Optional
Speaker Demographics						
ddpv:participants →	DDPV	Participant(s) information in the audio file, including speakers, interviewers, and other contributors.	Object	MANY	Manual / Automatic	Recommended
olac:role	OLAC	Participant role e.g. interviewer, interviewee	Text	ONE	Manual / Automatic	Recommended
olac:code	OLAC	Unique identifier for speaker	Text	ONE	Manual / Automatic	Recommended
sc:gender	Schema.org	Gender of the speaker	Text	ONE	Manual	Optional

ddpv:ageRange	DDPV	Age range of the speaker	Text	ONE	Manual	Optional
ddpv:dialectRegion	DDPV	Dialect, accent, or regional variety for the speaker	Text	ONE	Manual	Optional

Video Schema Attributes

Field Name	**Standard**	**Purpose**	**Expected Type**	**Cardinality**	**Entry Method**	**Requirement**
ebucore:hasTranscrip	EBUCore	Transcript or subtitle file for the audio or video	Text	ONE	Manual/ Automatic	Recommended
ebucore:duration	EBUCore	Length of individual media files	Number	ONE	Automatic	Recommended
ddpv:TotalDurationHours	DDPV	Total hours of content	Number	ONE	Automatic	Recommended
ebucore:bitDepth	EBUCore	How precisely each sample's amplitude is stored.	Number	ONE	Automatic	Optional
ebucore:width	EBUCore	Frame width in pixels	Number	ONE	Automatic	Recommended

ebucore:height	EBUCore	Frame height in pixels	Number	ONE	Automatic	Recommended
ebucore:frameRate	EBUCore	Frames per second	Number	ONE	Automatic	Recommended
ebucore:videoCodec	EBUCore	Video codec (e.g., 'H.264', 'H.265/HEVC', 'VP9', 'AV1')	Text	ONE	Automatic	Recommended
ebucore:bitDepth	EBUCore	Bits per color component	Integer	ONE	Automatic	Recommended
ddpv:equipmentType	DDPV	Recording device or microphone model	Text	ONE	Automatic	Optional
Speaker Demographics						
ddpv:participants →	DDPV	Participant(s) information in the audio file, including speakers, interviewers, and other contributors.	Object	MANY	Manual / Automatic	Recommended
olac:role	OLAC	Participant role e.g. interviewer, interviewee	Text	ONE	Manual / Automatic	Recommended

olac:code	OLAC	Unique identifier for speaker	Text	ONE	Manual / Automatic	Recommended
sc:gender	Schema.org	Gender of the speaker	Text	ONE	Manual	Optional
ddpv:ageRange	DDPV	Age range of the speaker	Text	ONE	Manual	Optional
ddpv:dialectRegion	DDPV	Dialect, accent, or regional variety for the speaker	Text	ONE	Manual	Optional

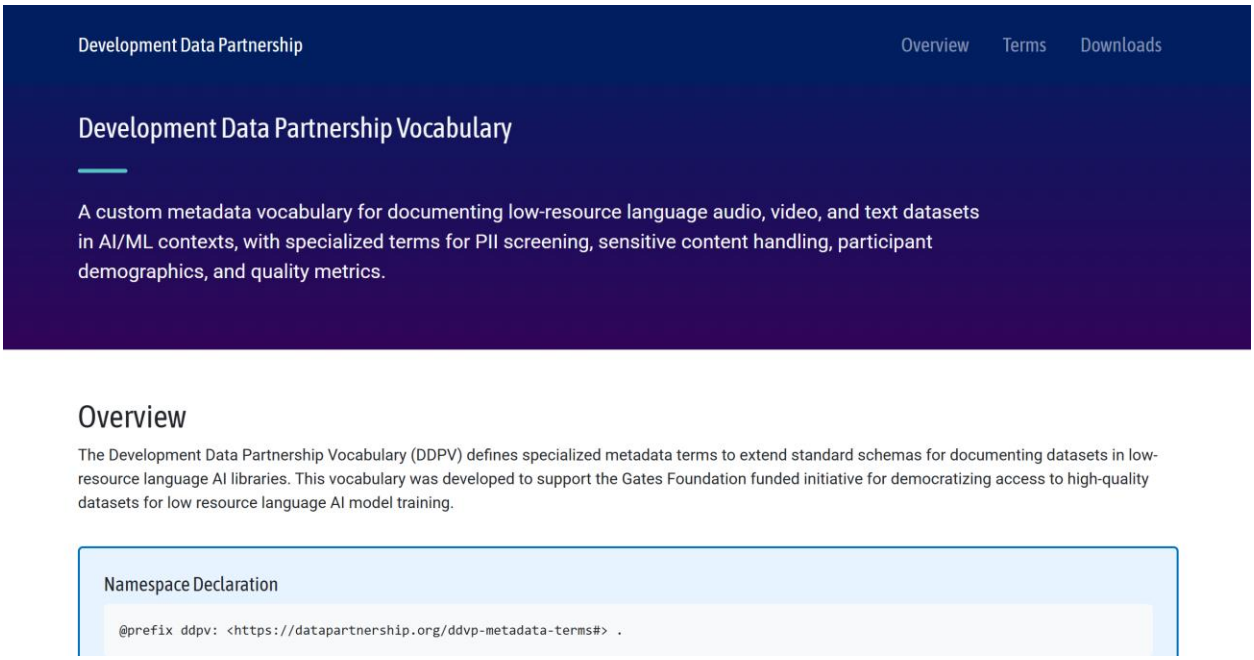
Metadata Collection and Implementation

As observed in the schemas above, metadata collection for this project employs a hybrid approach combining automated and manual methods. Most metadata will be extracted programmatically through automated processing pipelines that analyze datasets at scale. In some occasions, descriptive and contextual metadata, particularly for text documents, are collected through manual annotation using Chambo, a custom web-based metadata editor developed for this work.

Chambo supports three user roles: Admins, who manage users, configure metadata schemas, and assign files; Supervisors, who review, validate, and approve or modify metadata; and Taggers, who apply structured metadata to assigned documents. Core capabilities include role-based access and task assignment, configurable schemas, tag versioning and validation tracking, dashboards with search and audit trails, and notifications for assignments and review outcomes.

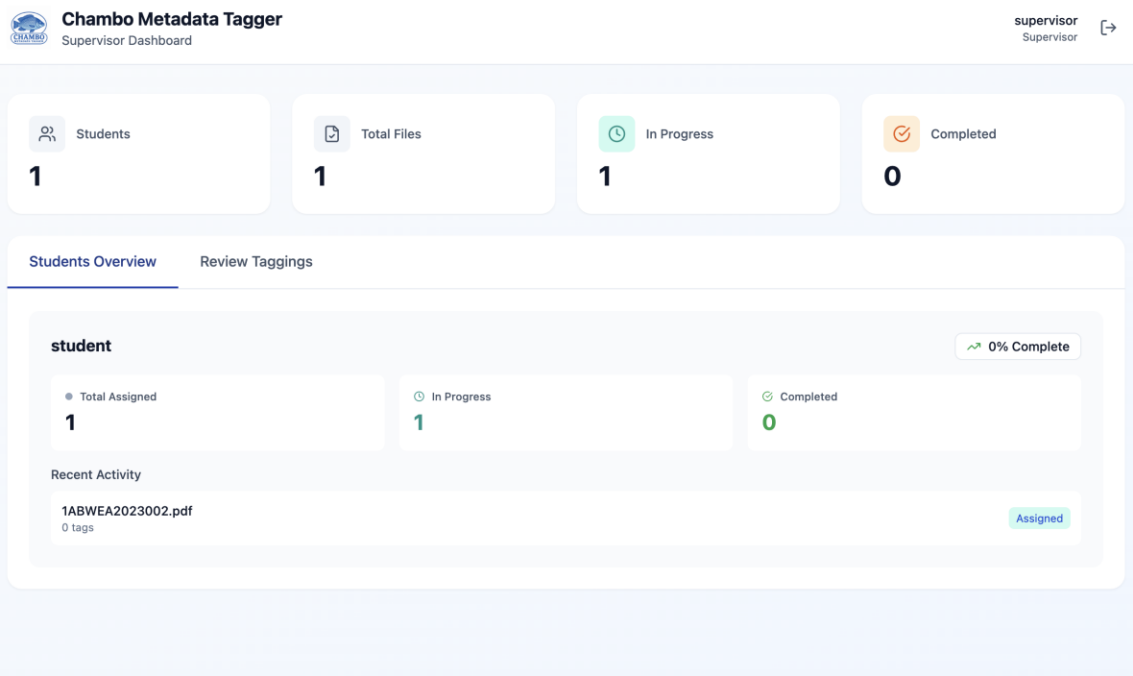
Annex 1: DDPV Web Page

The image below shows a screenshot of the DDPV webpage. The JSON-LD files for the schemas are available for download on this page and provide the full schema descriptions.

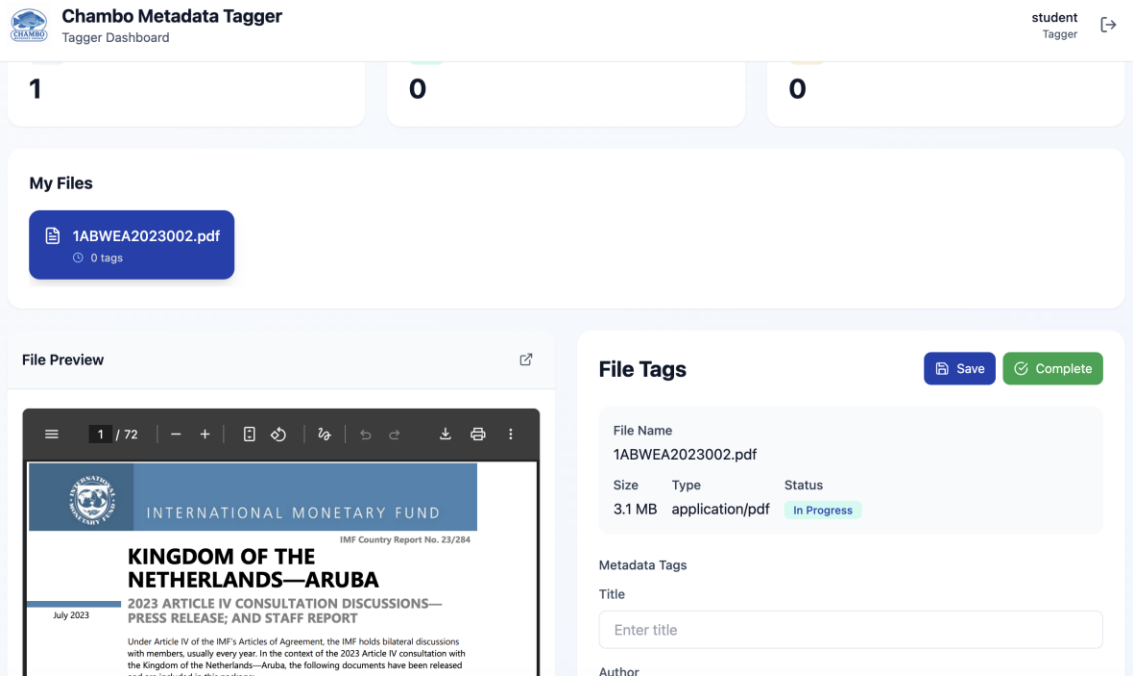


Annex 2: Metadata Tagging Tool

The image below shows the supervisor dashboard of the metadata editor.



The image below displays the student dashboard of the metadata editor.



References

[1] J. Greenberg, "Understanding metadata and metadata schemes," Cataloging & Classification Quarterly, vol. 40, no. 3–4, pp. 17–36, 2005.

[2] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, 160018, 2016.

[3] T. Gebru et al., "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021.

[4] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, "Dublin Core metadata for resource discovery," Internet Engineering Task Force RFC 2413, 1998.

[5] ISO, ISO 15836-1:2017 — Information and documentation — The Dublin Core metadata element set — Part 1: Core elements, International Organization for Standardization, 2017.

[6] R. Heery and M. Patel, "Application profiles: Mixing and matching metadata schemas," Ariadne, no. 25, 2000.

[7] T. R. Bruce and D. I. Hillmann, "The continuum of metadata quality: Defining, expressing, exploiting," D-Lib Magazine, vol. 10, no. 5, 2004.

[8] A. Hedstrom and C. A. Lee, "Significant properties of digital objects: Definitions, applications, implications," in Proceedings of the DLM-Forum, 2002.

[9] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, and Y. Bengio, "Improving reproducibility in machine learning research," Journal of Machine Learning Research, vol. 22, no. 164, pp. 1–20, 2021.

[10] K. Coyle and T. Baker, Guidelines for Dublin Core™ Application Profiles, Dublin Core Metadata Initiative (DCMI) Recommended Resource, May 18, 2009.

[11] M. Nilsson, T. Baker, and P. Johnston, "The Singapore framework for Dublin Core application profiles," in Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008.

[12] MLCommons, Croissant: A metadata standard for machine learning datasets, MLCommons, Mar. 2024.

[13] R. Guha, D. Brickley, and S. Macbeth, "Schema.org: Evolution of structured data on the web," Communications of the ACM, vol. 59, no. 2, pp. 44–51, 2016.

[14] W3C, Data Catalog Vocabulary (DCAT) – Version 2, W3C Recommendation, 2020.

[15] W3C, Data Quality Vocabulary (DQV), W3C Working Group Note, 2016.

[16] W3C, PROV-O: The PROV Ontology, W3C Recommendation, 2013.

[17] European Broadcasting Union (EBU), EBUCore Metadata Set, Technical Specification, 2018.

[18] S. Bird and G. Simons, "OLAC: Accessing the world's language resources," Proceedings of the Workshop on Web-Based Language Documentation, 2003.

[19] B. D'Arcus and F. Giasson, Bibliographic Ontology Specification (BIBO), 2009.

[20] M. Vardigan, P. Heus, and J. Thomas, "Data Documentation Initiative: Toward a standard for the social sciences," in Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008.

[21] W3C, RDF Schema 1.1, W3C Recommendation, 2014.