

(Free) tools for data documentation and dissemination

(using DDI Codebook for microdata)

November 2022

Contacts:

Mehmood Asghar: masghar@worldbank.org

Olivier Dupriez: odupriez@worldbank.org

Aivin Solatorio: asolatorio@worldbank.org

Components

- 1 Guide** on the use of metadata standards and schemas
 - Advocacy for the use of metadata standards/schemas
 - Technical guidelines
- 2 Cataloguing application: NADA (National Data Archive)**
 - Open source
 - Multi-standard, including DDI Codebook
- 3 Metadata Editor**
 - To replace the Nesstar Publisher; multi-standard including DDI Codebook
 - In testing phase
- 4 Research/exploratory work** on improved data discoverability
 - Use of machine learning for semantic searchability, recommender system, metadata augmentation

A set of standards for multiple data types



DDI CODEBOOK
2.5 FOR
MICRODATA



ISO19139/19115/19110
FOR GEOGRAPHIC
DATASETS



DUBLIN CORE/
MARC21/BIBTEX FOR
DOCUMENTS



DUBLIN CORE/IPTC
FOR IMAGES



CUSTOM SCHEMAS FOR
INDICATORS, TABLES,
VIDEOS, REPRODUCIBLE
SCRIPTS



ALL TYPES: MAPPING
TO SCHEMA.ORG
FOR SEARCH ENGINE
OPTIMIZATION

Guidelines on standards and schemas

Metadata Schema Guide

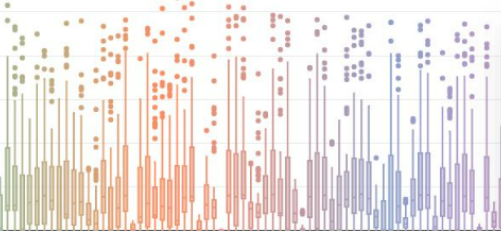
- Preface
- Introduction
- I GET STARTED
 - 1 The benefits of rich, structured met...
 - 2 Metadata formats and tools
 - 3 Data cataloging
- II STANDARDS AND SCHEMAS
 - 4 Documents
 - 5 Microdata
 - 6 Geographic data and services
 - 7 Indicators and time series
 - 8 Statistical tables
- 9 Images
- 10 Audio
- 11 Videos
- 12 Research projects and scripts
- 13 External resources
- ANNEXES
 - Annex 1: References and links
 - Annex 2: Mapping standards and sch...
 - Annex 3: Mapping the microdata sch...
 - Annex 4: Mapping the geographic sc...

[DRAFT] Metadata Standards and Schemas for Improved Data Discoverability and Usability

Olivier Dupriez and Mehmood Asghar

2022-11-15

Preface



The volume and the diversity of data made available to the research community is growing rapidly. However, many valuable datasets remain largely under-exploited. To be used more effectively, data must be easier to find, understand, access, and analyze. The documentation and quality of the associated metadata must be improved. This document is a guide on the

Search

- CATALOG ADMINISTRATION
- COLLECTIONS >
- DATASETS >
- EXTERNAL RESOURCES >
- SCRIPTS >
- SURVEY >
 - POST Create survey
 - POST Update survey
 - GET List data files
 - POST Create data file
 - DELETE Delete datafile
 - DELETE Delete variable
 - GET List variables

Create survey

Create a new survey

PARAMETERS

Path Parameters

- IDNo string <string> Required
Dataset IDNo

REQUEST BODY

- repositoryid Collection ID that owns the survey (string)
Abbreviation for the collection that owns this survey.
- access_policy Data access policy (string)
Default: "data_na"

POST /datasets/create/survey...

REQUEST SAMPLES

```
{
  "repositoryid": "string",
  "access_policy": "data_na",
  "published": 0,
  "overwrite": "no",
  "doc_desc": {
    "title": "string",
    "idno": "string",
    "producers": [ - ],
    "prod_date": "string",
    "version_statement": { - }
  },
  "study_desc": {
    "title_statement": { - },
    "authoring_entity": [ - ],
    "oth_id": [ - ],
    "production_statement": { - },
    "distribution_statement": { - },
    "series_statement": { - },
    "version_statement": { - },
    "bib_citation": "string",
    "bib_citation_format": "string",
    "holdings": [ - ],
    "study_notes": "string",
    "study_authorization": { - },
    "study_info": { - },
    "study_development": { - },
    "method": { - },
    "data_access": { - }
  },
  "data_files": [
    { - }
  ],
  "variables": [
    { - }
  ],
  "variable_groups": [
    { - }
  ]
}
```

Metadata Schema Guide


- Preface
- Introduction
- I GET STARTED
 - 1 The benefits of rich, structured met...
 - 2 Metadata formats and tools
 - 3 Data cataloging
- II STANDARDS AND SCHEMAS
 - 4 Documents
 - 5 Microdata
 - 5.1 Definition of microdata
 - 5.2 The Data Documentation Initiative
 - 5.3 Some practical considerations
 - 5.4 Schema description: DDI-Cod...
 - 5.5 Generating and publishing DD...
 - 6 Geographic data and services
 - 7 Indicators and time series
 - 8 Statistical tables

Chapter 5 Microdata

5.1 Definition of microdata

When surveys or censuses are conducted, or when administrative data are recorded, information is collected on each unit of observation. The unit of observation can be a person, a household, a firm, an agricultural holding, a facility, or other. Microdata are the data files resulting from these data collection activities, which contain the unit-level information (as opposed to aggregated data in the form of counts, means, or other). Information on each unit is stored in *variables*, which can be of different types (e.g. numeric or alphanumeric, discrete or continuous). These variables may contain data reported by the

Chapter 5 Microdata



5.1 Definition of microdata

When surveys or censuses are conducted, or when administrative data are recorded, information is collected on each unit of observation. The unit of observation can be a person, a household, a firm, an agricultural holding, a facility, or other. Microdata are the data files resulting from these data collection activities, which contain the unit-level information (as opposed to aggregated data in the form of counts, means, or other). Information on each unit is stored in *variables*, which can be of different types (e.g. numeric or alphanumeric, discrete or continuous). These variables may contain data reported by the

<https://mah0001.github.io/schema-guide/>

<https://ihsn.github.io/nada-api-redoc/catalog-admin/#tag/Survey>

NADA: a multi-standard cataloguing tool

- Open source
 - <https://github.com/ihsn/nada>
 - <https://ihsn.github.io/nada-documentation/>
- Technologies: PHP; SQL (metadata); mongoDB (data); Solr (optional)
- Widgets for flexible additions (e.g., embed visualizations or data grids in catalog pages using JS libraries of your choice)
- R package and Python library for automation of tasks
 - To generate, harvest, publish, edit, augment, extract, import/export metadata
- Internationalization: embedded translation tool

2

NADA: multiple data types / standards

- Multi-standard (for multiple **data** types and **reproducible scripts**)
- Data organized by type and (optional) by collection
 - E.g., thematic collections, and/or collections by data producer



Keywords... Search

All Microdata Geospatial Time series Tables Documents Images Videos Scripts

10 2 10 18 7 6 6 1

Years Countries Collections Microdata Access Data Type

Showing 1-15 of 60 Study view Variable view Relevance

Advancing Gender Equality through Household Surveys: Living Standards Measurement Study-Plus (LSMS+)
2021
ID: VDO_005 Last modified: Apr 09, 2022 Views: 85

Age dependency ratio (% of working-age population)
Aruba, Africa Eastern and Southern, Afghanistan...and 239 more, 1960-2020
ID: SP.POP.DPND Last modified: Oct 29, 2021 Views: 562

Bangladesh, Outline of camps of Rohingya refugees in Cox's Bazar, January 2021
2021
Collection: South Asia

2

NADA: rich, structured metadata

- Rich metadata
- Flexible data access control
- Can federate the catalog administration
 - Sub collections
 - Fine-grained roles/permissions system
- Embedded SEO (Google data structure / schema.org)
- Metadata accessible via API
- R package for automation of tasks
- Responsive design (bootstrap4)

The screenshot shows a web page for the 'Demographic and Health Survey 2009' in Maldives. The page is part of a 'Central Data Catalog' and includes a 'GET MICRODATA' button. It displays various metadata fields such as Reference ID, Producer(s), Collections, and Metadata. A sidebar on the right shows statistics like 'CREATED ON', 'LAST MODIFIED', 'PAGE VIEWS', and 'DOWNLOADS'. The main content area is divided into sections: 'STUDY DESCRIPTION', 'DATA DESCRIPTION', 'DOWNLOADS', 'GET MICRODATA', and 'RELATED PUBLICATIONS'. The 'Identification' section is currently active, showing details like 'SURVEY ID NUMBER', 'TITLE', 'COUNTRY', 'STUDY TYPE', 'SERIES INFORMATION', and 'ABSTRACT'.

Home / Central Data Catalog / MDV_2009_DHS_v01_M

Demographic and Health Survey 2009
Maldives, 2009 [GET MICRODATA](#)

Reference ID MDV_2009_DHS_v01_M

Producer(s) Ministry of Health and Family (MoHF)

Collections [South Asia](#)

Metadata [Documentation in PDF](#) [DDI/XML](#) [JSON](#)

CREATED ON
Sep 13, 2021

LAST MODIFIED
Sep 13, 2021

PAGE VIEWS
730

DOWNLOADS
5

STUDY DESCRIPTION DATA DESCRIPTION DOWNLOADS [GET MICRODATA](#) RELATED PUBLICATIONS

Identification

SURVEY ID NUMBER
MDV_2009_DHS_v01_M

TITLE
Demographic and Health Survey 2009

COUNTRY

Name	Country code
Maldives	MDV

STUDY TYPE
Demographic and Health Survey (standard) - DHS V

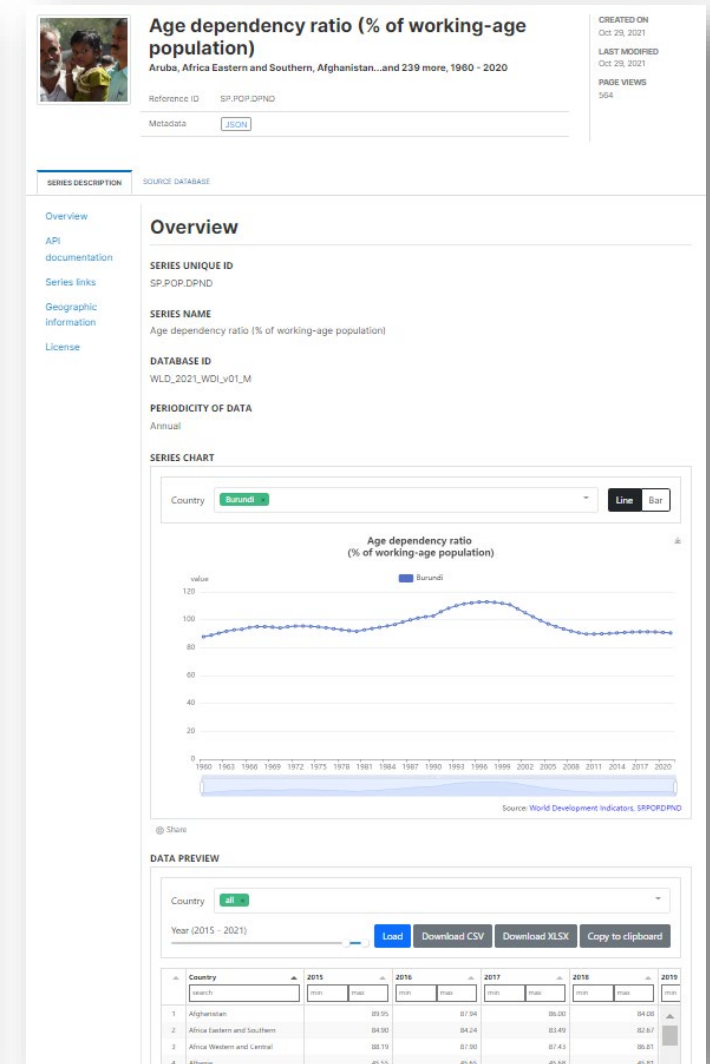
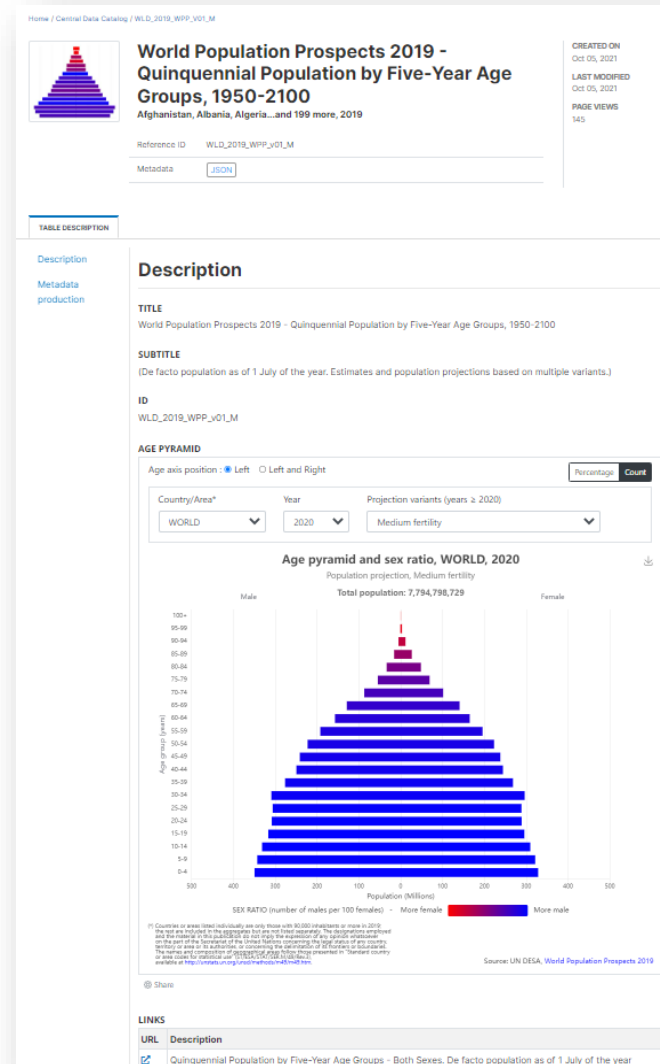
SERIES INFORMATION
The DHS 2009 is the first Demographic and Health Survey conducted in the Maldives.

ABSTRACT
The 2009 MDHS was designed to provide data to monitor the population and health situation in Maldives. Specifically, the MDHS collected information on fertility levels and preferences, marriage, sexual activity, knowledge and use of family planning methods, breastfeeding practices, nutrition status of women and young children, childhood mortality, maternal

NADA: embed visualizations and more

Embed visualizations, data grids, maps, etc. created with other applications, e.g., eCharts

- Mostly for indicators
- Requires data accessible via API (NADA API or external API)
- Can build your own tools using NADA as back-end



NADA: variable-level search and comparison

For microdata: variable-level search and comparison tools

drinking water × Search

Years ▼ Showing 1-15 of 19 Study view Variable view Relevance ▼

Countries ▼

Authoring Entity ▼ drinking water × Reset search

Collections ▼

Microdata Access ▼

Compare

- hv201 - Source of drinking water**
n/a
Demographic and Health Survey 2009
Maldives - MDV_2009_DHS_v01_M
- hv202 - Source of non-drinking water**
n/a
Demographic and Health Survey 2009
Maldives - MDV_2009_DHS_v01_M
- sh108 - Experienced a shortage of drinking water**
n/a
Demographic and Health Survey 2009
Maldives - MDV_2009_DHS_v01_M
- v113 - Source of drinking water**
n/a

Compare variables 2 ▲

Compare Variables

Refresh Clear Download variables as PDF CSV JSON ↗

WS1 × Remove

Kazakhstan
Multiple Indicator Cluster Survey 2006

Main source of drinking water (WS1)
Data file: [hh](#)

Overview

Valid: 14564	Type: Discrete
Valid (weighted): 14564.001	Decimal: 0
Invalid: 436	Start: 42
Invalid (weighted): 0	End: 43
Minimum: 11	Width: 2
Maximum: 96	Range: 11 - 99
	Format: Numeric
	Weighted variable: V896

Questions and instructions

LITERAL QUESTION
What is the main source of drinking water for members of your household?

CATEGORIES

Value	Category	Cases	Wt
11	Piped into dwelling	6932	77%

hv201 × Remove

Maldives
Demographic and Health Survey 2009

Source of drinking water (hv201)
Data file: [RECH2](#)

Overview

Type: Discrete
Decimal: 0
Start: 6
End: 7
Width: 2
Range: 10 - 99
Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
10	PIPED WATER
11	Piped into dwelling
12	Piped to yard/plot
13	Public tap/standpipe

v113 × Remove

Maldives
Demographic and Health Survey 2009

Source of drinking water (v113)
Data file: [REC11](#)

Overview

Type: Discrete
Decimal: 0
Start: 21
End: 22
Width: 2
Range: 10 - 99
Format: Numeric

Questions and instructions

CATEGORIES

Value	Category
10	PIPED WATER
11	Piped into dwelling
12	Piped to yard/plot
13	Public tap/standpipe

2

NADA: data deposit system

Data deposit system for controlled acquisition of data/metadata

PENDING TASKS

STUDY DESCRIPTION
▲ Fill 2 mandatory field(s)


UPLOAD FILES
▲ No files uploaded

CITATIONS
▲ No citations

MY PROJECTS

- [Test survey deposit Olivier \(draft\)](#)

1 **Project Information**

2 **Study Description** 
(You are here)

3 **Data files and other Resources**

4 **Citations (optional)**

5 **Review and Submit**

Study description

Please complete the fields in each of the sections below. Providing detailed information here will speed up the process of publishing the study. It also makes it easier for users of the data to find the information they need and thus lessen the need for users to contact the data producer for clarification. Only three fields are mandatory for the submission process. If time or information available does not allow for the completion of all fields then we request that at least the mandatory and recommended fields be completed.

Settings ▾ [Import Metadata](#) [Expand All](#) [Collapse All](#)

Identification

* Title *Required*

Subtitle

Abbreviation

Study Type

NADA: Management using R or Python (APIs)

```

R
library(nadar)

# -----
my_keys <- read.csv("C:/confidential/my_API_keys.csv", header=F, stringsAsFactors=F)
set_api_key("my_keys[1,1]")
set_api_url("https://.../index.php/api/")
set_api_verbos(FALSE)
# -----

setwd("C:/my_folder")
doc_file <- "WB_PRNP_9412_Food_Crises.pdf"

id <- "WB_WPS9412"

thumb_file <- gsub(".pdf", ".jpg", doc_file)
capture_pdf_cover(doc_file) # Capture cover page for use as thumbnail

example_1 <- list(

  document_description = list(

    title_statement = list(idno = id, title = "Predicting Food Crises"),

    date_published = "2020-09",

    authors = list(
      list(last_name = "Andrée", first_name = "Bo Pieter Johannes",
           affiliation = "World Bank",
           author_id = list(list(type = "ORCID", id = "0000-0002-8007-5007"))),
      list(last_name = "Chamorro", first_name = "Andres",
           affiliation = "World Bank"),
      list(last_name = "Kraay", first_name = "Aart",
           affiliation = "World Bank"),
      list(last_name = "Spencer", first_name = "Phoebe",
           affiliation = "World Bank"),
      list(last_name = "Wang", first_name = "Dieter",
           affiliation = "World Bank",
           author_id = list(list(type = "ORCID", id = "0000-0003-1287-332X")))
    ),

    journal = list(
      name = "World Bank Policy Research Working Paper",
      number = "9412",
      publisher = "World Bank",
    ),


    ref_country = list(
      list(name="Afghanistan", code="AFG"),
      list(name="Burkina Faso", code="BFA"),
      list(name="Chad", code="TCD")
    )
  )

```

Generate
metadata
using R or
Python and
publish it in
NADA.

Use of NADA
API allows
automation of
many tasks
(scraping or
harvesting,
transforming,
checking,
documenting,
publishing)

Home / Central Data Catalog / WB_WPS9412



Predicting Food Crises

Afghanistan, Burkina Faso, Chad...and 18 more, 2020

Reference ID: WB_WPS9412

Producer(s): Bo Pieter Johannes Andrée, Andres Chamorro, Aart Kraay, Phoebe Spencer, Dieter Wang

Collections:

Metadata:

CREATED ON
Feb 17, 2022

LAST MODIFIED
Feb 17, 2022

DOCUMENT DESCRIPTION

Description

Authoring information

Reproducibility

Description

UNIQUE USER DEFINED ID
WB_WPS9412

TITLE
Predicting Food Crises

DOWNLOADS
Predicting Food Crises [Download](#)

JOURNAL NAME
World Bank Policy Research Working Paper

DATE PUBLISHED

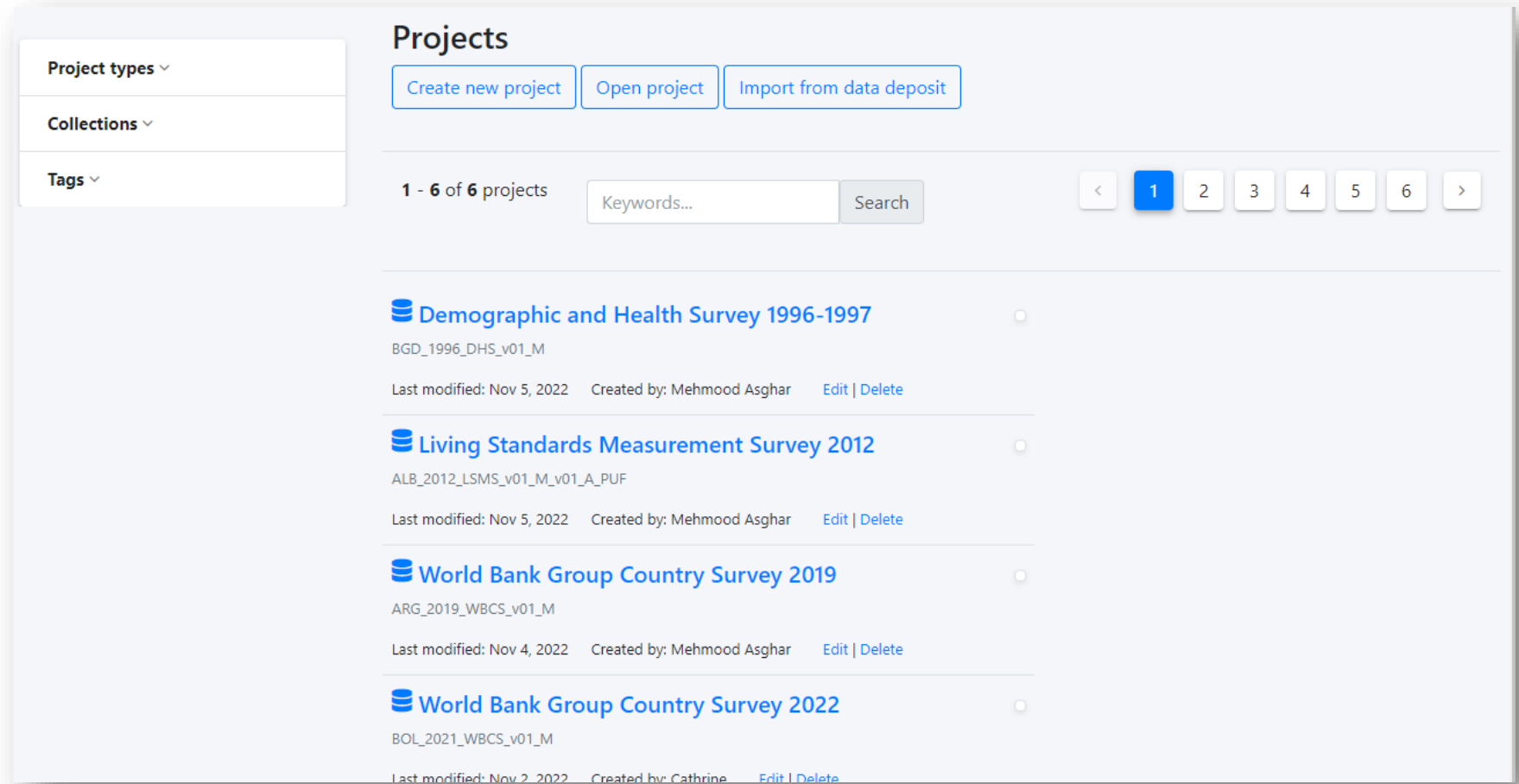
Full examples available in
<https://mah0001.github.io/schema-guide/chapter04.html>

Metadata Editor (DDI Codebook for microdata)

- To replace the Nesstar Publisher (and more)
- Multi-standards; can accommodate user-defined standards/schemas
 - Easy to upgrade when new versions of standards like DDI Codebook are released
- Uses R (haven package) to import microdata files/generate summary statistics
 - Easy to upgrade when new data file formats are released
- Multi-platform; stand-alone or server application
- In final testing phase (as of Dec. 2022); public release around March 2023
- In future version: add metadata augmentation utilities using plugins
 - Connect to machine learning APIs for keyword/topic extraction, classification and tagging, word embeddings; image labeling; speech-to-text for videos, translation, ...)

Metadata Editor – Home page

- “Project type” can be microdata, geospatial, image, document, table, indicator, etc.
- Access to projects controlled by user login
- Projects can be:
 - Private
 - Shared with a selected group
 - Shared with all logged-in curators



The screenshot displays the 'Projects' section of the Metadata Editor. On the left, there are three filter menus: 'Project types', 'Collections', and 'Tags'. The main content area features a header with three buttons: 'Create new project', 'Open project', and 'Import from data deposit'. Below this is a search bar with the text '1 - 6 of 6 projects' and a search input field labeled 'Keywords...' with a 'Search' button. A pagination control shows page 1 selected. The project list includes:

Project Name	ID	Last modified	Created by	Actions
Demographic and Health Survey 1996-1997	BGD_1996_DHS_v01_M	Nov 5, 2022	Mehmood Asghar	Edit Delete
Living Standards Measurement Survey 2012	ALB_2012_LSMS_v01_M_v01_A_PUF	Nov 5, 2022	Mehmood Asghar	Edit Delete
World Bank Group Country Survey 2019	ARG_2019_WBCS_v01_M	Nov 4, 2022	Mehmood Asghar	Edit Delete
World Bank Group Country Survey 2022	BOL_2021_WBCS_v01_M	Nov 2, 2022	Cathrine	Edit Delete

3

Metadata Editor – Study-level metadata page

Process: like in
Nesstar Publisher

1. Select a template
2. Import data (for microdata)
3. Add metadata and external resources
4. Augment metadata (in future version)
5. Save/export/publish
 - Project saved as a ZIP file that includes the data, metadata, external resources, study thumbnail, and the template.

The screenshot shows the 'Metadata Editor' interface for a study titled 'World Bank Group Country Survey 2019'. The interface is divided into a left sidebar and a main content area. The sidebar contains a navigation menu with categories like 'Home', 'Document description', 'Study description', and 'Identification'. The 'Identification' category is expanded, showing fields for IDNO, Title, Subtitle, Alternate title, Translated title, Series type, and Series name. The main content area displays the 'Identification' section with input fields for these fields. The 'IDNO' field contains 'COL_2019_WBCS_v01_M'. The 'Title' field contains 'World Bank Group Country Survey 2019'. The 'Series name' field contains a detailed description of the survey program.

World Bank Group Country Survey 2019
survey - COL_2019_WBCS_v01_M

Templates Metadata Publish

Identification

IDNO *

Title *

Subtitle

Alternate title

Translated title

Series type

Series name

The World Bank Group Country Opinion Survey Program systematically measures and tracks the perceptions of the World Bank's clients, partners, and other stakeholders across the globe in client countries. The Public Opinion Research Group surveyed 29 countries in FY2012 (July 2011-June 2012), 41 countries in FY2013 (July 2012-June 2013), 42 countries in FY2014 (July 2013-June 2014), 35 countries in FY2015 (July 2014-June 2015), 45 countries in FY2016 (July 2015-June 2016), 35 countries in FY2017 (July 2016-June 2017), and 39 countries in FY2018 (July 2017-June 2018). In FY2019, surveys were conducted in 42 countries. Nearly all of the World Bank Group's client countries are surveyed in every three year cycle.

3

Metadata Editor – Variable-level metadata page

Imports/ exports microdata files and generates summary statistics using R (haven package)



Data files

40 files

[Create file](#) [Import file](#)

File ID	File name	Variables			
F1	RECH0	34			
F2	RECH1	20			
F3	RECH2	23			
F4	RECH3	14			
F5	RECH4	4			
F6	REC01	37			

World Bank Group Country Survey 2019
survey - COL_2019_WBCS_v01_M

Variables 269

Variable ID	Label	Scale
V1	id	id
V2	method	Survey completion method
V3	a1	When you think about the future in Colombia, are you ... ?
V4	a2	Do you think that economic opportunity for citizens in Colombia is ... ?
V5	a3_1	Urban development
V6	a3_2	Energy
V7	a3_3	Water and sanitation
V8	a3_4	Balanced territorial development
V9	a3_5	Job creation/employment
V10	a3_6	Private sector development/entrepreneurship
V11	a3_7	Education
V12	a3_8	Public sector governance/reform
V13	a3_9	Global/regional integration

Documentation a2 - Do you think that economic opportunity for citizen

STATISTICS WEIGHTS DOCUMENTATION JSON

Frequencies

Value	Label	Cases	Weighted
1	Increasing	52	
2	Decreasing	74	
3	Staying about the same	74	

Summary statistics

Valid	200
Invalid	
min	1
max	3

Categories

Value	Label	
1	Increasing	
2	Decreasing	
3	Staying about the same	

Variable information

Interval type
Discrete

Decimal points
0

Format
Numeric

3

Metadata Editor – Data preview page

Imports/ exports data files and generates summary statistics using R (haven package)



Metadata Editor

World Bank Group Country Survey 2019
survey - ARG_2019_WBCS_v01_M

Templates Metadata Publish

Editor

- Home
- Document description
- Study description
- Tags
- Data files
 - argentina_cos_fy19_datafile
 - Variables
 - Data**
- External resources

Data

Showing records 1 - 50 of 375

Import Data Export Data

#	id	method	a1	a2	a3_1	a3_2	a3_3	a3_4	a3_5	a3_6	a3_7	a3_8	a3_9	a3_10	a3_11	a3_12	a3_13	a3_14
1	101	Online	Somewhat pessimistic	Decreasing	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	102	Online	Somewhat pessimistic	Decreasing	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3	103	Online	Somewhat pessimistic	Decreasing	0	0	0	1	0	0	0	0	0	0	0	0	0	1
4	104	Online	Very optimistic	Staying about the same	0	0	0	1	0	0	0	0	0	0	0	1	0	0
5	105	Online	Somewhat optimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
6	106	Online	Somewhat pessimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	107	Online	Somewhat optimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
8	108	Online	Somewhat pessimistic	Staying about the same	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	109	Online	Somewhat pessimistic	Staying about the same	0	0	0	1	0	0	0	0	0	0	0	1	1	0
10	110	Online	Somewhat optimistic	Decreasing	0	0	0	1	0	0	0	0	0	1	0	0	0	0
11	111	Online	Somewhat pessimistic	Decreasing	1	0	0	1	0	0	0	0	0	0	0	0	0	0
12	112	Online	Somewhat pessimistic	Decreasing	0	0	0	0	0	0	0	0	0	0	0	1	0	0
13	113	Online	Somewhat optimistic	Decreasing	0	0	1	0	0	0	0	0	0	0	0	0	0	0
14	114	Online	Somewhat pessimistic	Staying about the same	0	0	0	0	1	0	0	0	0	0	0	0	0	0
15	115	Online	Somewhat optimistic	Staying about the same	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	116	Online	Not sure	Decreasing	0	0	0	0	0	0	1	0	0	0	0	0	0	0

3

Metadata Editor – Document other data types

Works similarly for other data types, with their own standard or schema, e.g, using elements from DCMI and MARC21 to document a **publication**.

The screenshot displays the Metadata Editor interface. On the left is a dark sidebar with a navigation menu. The main area is a light gray form titled 'TODO document -'. The form is organized into sections: 'Title statement' (containing 'IDNo' and 'Title' fields), 'Subtitle', 'Alternate title', and 'Abbreviated Title'. The sidebar menu includes categories like 'Home', 'Document', 'Title statement', 'Authors and Contributors', 'Date information', 'Identifiers', 'Description', 'Tags', 'Metadata Description', and 'External resources'. The top of the form has a header with 'Templates', 'Metadata', and 'Publish' options.

Research/exploratory work on NLP

- Objective: **improved discoverability**; exploit natural language processing (NLP) models to build a semantic search and recommender system for data cataloguing applications (NADA and others)
- Also: build tools for automated metadata enhancement to be embedded in metadata editors
- Models/tools already tested with satisfactory results, but still needs tuning and tools development (open APIs)
- Exploratory work available in nlp4dev platform (pwd = “nlpexplorer”)
<https://www.nlp4dev.org/>

Research/exploratory work on NLP

Testing a semantic search and recommender system

Example:

User can submit a PDF document as a query

→ Document is processed and analyzed

→ Closest matches are returned (ranked by semantic closeness)

The screenshot displays the NLP4Dev website interface. At the top, the header includes the logo "NLP4Dev" with the tagline "Natural Language Processing for knowledge discovery" and navigation links for Home, Search, Explore, Methods & Tools, API, and About. Below the header is a search bar with a magnifying glass icon and the text "Search". A red arrow points to the search input field, which contains the filename "sachsmalariafeb02.pdf". To the right of the search bar is a blue "Search" button. Below the search bar, there are radio buttons for "Keyword search" and "Semantic search", with "Semantic search" selected. A note below the radio buttons says "You can also try filtering by topic composition".

On the left side of the search results, there are three filter panels: "Year" with "From" and "To" dropdown menus, "Document type" with a dropdown arrow, and "Source" with a dropdown arrow. The main search results area shows "Showing 1-10 of 409949 documents". The first result is titled "Disease and Mortality in Sub-Saharan Africa : second edition". The author list includes Florence Baingana, Eduard R. Bos, Richard G. A. Feacham, Karen J. Hofman, Dean T. Jamison, Makgoba, and Khama O. Rogo. The source is listed as "WB, Rank: 1". There are links for "Metadata", "Topics", and "Related documents graph". Below the main result, there is a "Related documents" section with three cards: "Disease and mortality in sub-Saharan Africa, Ghana, 1991", "The health of adults in the developing world, 1992", and "Reproductive, maternal health, World, 2016".

Research/exploratory work on NLP

- The closest **datasets** available in the data catalog are also returned (based on semantic closeness between the query/PDF document and the metadata available for each dataset)
- Could also use geographic coverage, year, and other criteria for ranking results)
- Requires rich and augmented metadata to return relevant results

NLP4Dev Natural Language Processing for knowledge discovery

Home Search Explore Methods & Tools API

Disease and Mortality in Sub-Saharan Africa : second edition

Author(s): Florence [editor] Baingana, Eduard R. [editor] Bos, Richard G. A [editor] Feacham, Karen J. [editor] Hofman, Dean T. [editor] Jamison, Malegapuru [editor] Makgoba, Khama O. [editor] Rogo

Since the publication of the first edition of "Disease and Mortality in Sub-Saharan Africa" (report no. 9784 (1991)), many new sources of health and demographic information have become available, including data on trends in HIV infection from antenatal clinic surveillance sites, the first set of African life tables from a growing number of demographic surveillance sites, injury statistics from a small number of i...

[read more](#)

Africa, 2006
Category: Publications and Reports
Source: WB
Open in: World Bank Documents and Reports
Created on: May 22, 2006 Last modified: Feb 24, 2022 Views: 0
Metadata [JSON](#)

Metadata View document Related documents **Related data** Related documents graph

Related World Development Indicators

Cause of death, by non-communicable diseases (% of total)

Cause of death refers to the share of all deaths for all ages by underlying causes. Non-communicable diseases include cancer, diabetes mellitus, cardiovascular diseases, digestive diseases, skin diseases, musculoskeletal diseases, and congenital anomalies.

[Link to data](#)
[Link to metadata](#)

1

2

3

4

Ongoing/planned activities, and collaboration

- Addition of multiple new features in NADA and Metadata Editor
- Exploratory work on use of machine learning for better data discoverability
 - Semantic search; D3 indexing for geographic data; query parsers; improved search results ranking; and more.
 - Objective: develop smarter search algorithms and a recommender system for data catalogs, and metadata enhancement tools (with open-source software / open APIs)
- Collaboration welcome in multiple areas, e.g.:
 - Review/improvement of our custom metadata schemas
 - Development and/or testing of software (Metadata Editor, NADA, R package, PyNADA)
 - Production of technical documentation and training materials for Metadata Editor/NADA
 - Translation of Metadata Editor and related documentation
 - Design new features for Metadata Editor and NADA
 - Support to resource-constrained organizations (in low-income countries)