



KENYA AFFORDABLE HOUSING DATA PROJECT

Developing indicator metadata

Agenda



- 1) About this workshop
- 2) Overview of the data pipeline
- 3) Developing indicator metadata
- 4) Automating indicator population

About this workshop

What this workshop aims to achieve

- Develop sufficiently detailed indicator metadata

Why do we need metadata? What distinguishes good metadata from bad? How should I go about developing my own metadata? This workshop will help participants to navigate some of the core uncertainties of the modern data pipeline

- Familiarization with the downstream data pipeline

Developing metadata and collection templates is only the first stage of the data pipeline. The last part of the workshop considers what happens next and the skills required to move further down the data pipeline

Key outcomes

Following the workshop, we want the KNBS team to

- Develop and finalize the metadata for the **priority** KNBS indicators
- Collect the data for the **priority** KNBS indicators
- Continue working on the metadata and data collection for the remaining KNBS indicators
- Start to feel comfortable about revising/updating/appending metadata going forward

PLEASE:

- Ask questions, we've been down this road before!
- Metadata may not be the most exciting thing in the world but future you says thanks!



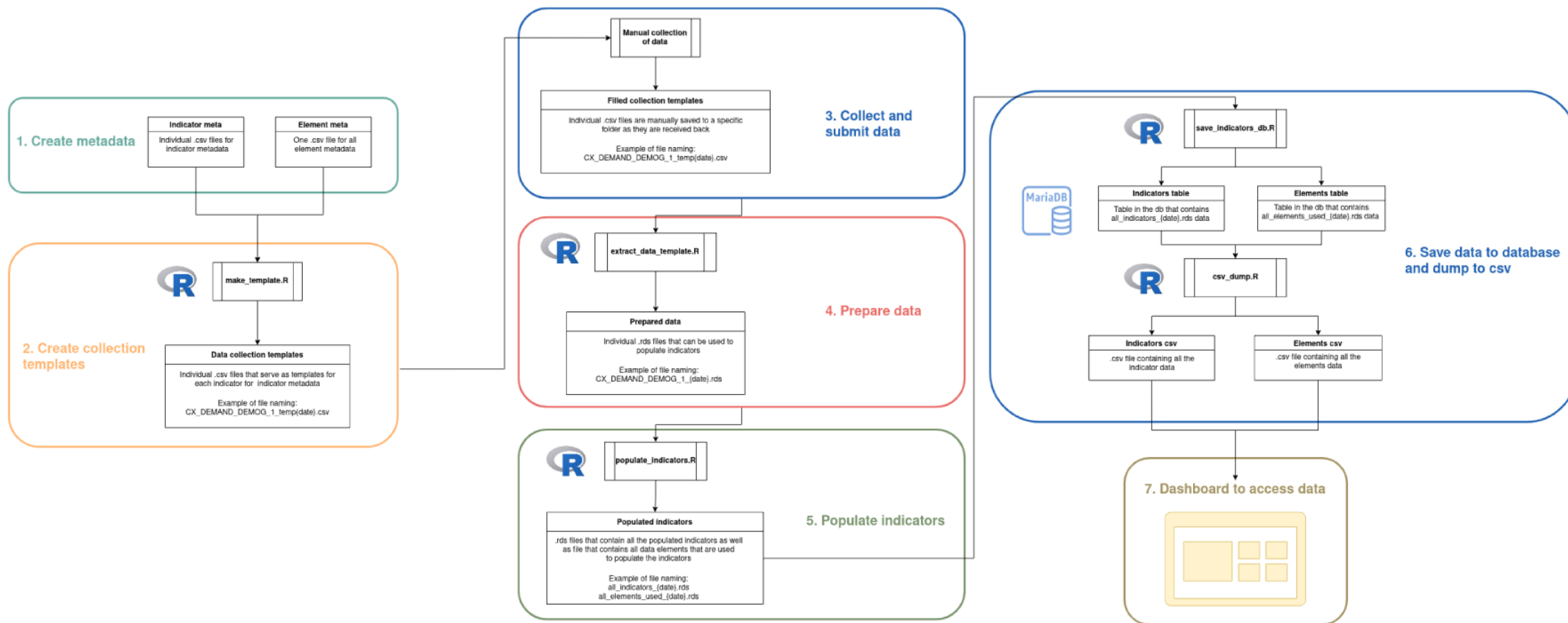
Asking for assistance



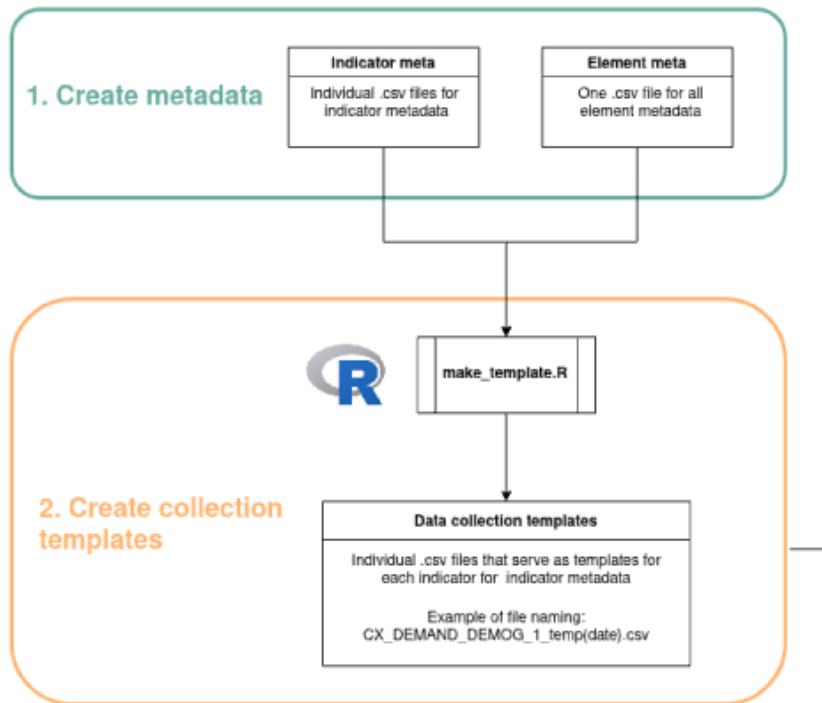
- Please raise your if you have a question and we'll get to you as soon as we can
- The is also being monitored for questions and the session will be recorded
- Please remain when not asking a question

Automated data pipeline

The full data pipeline

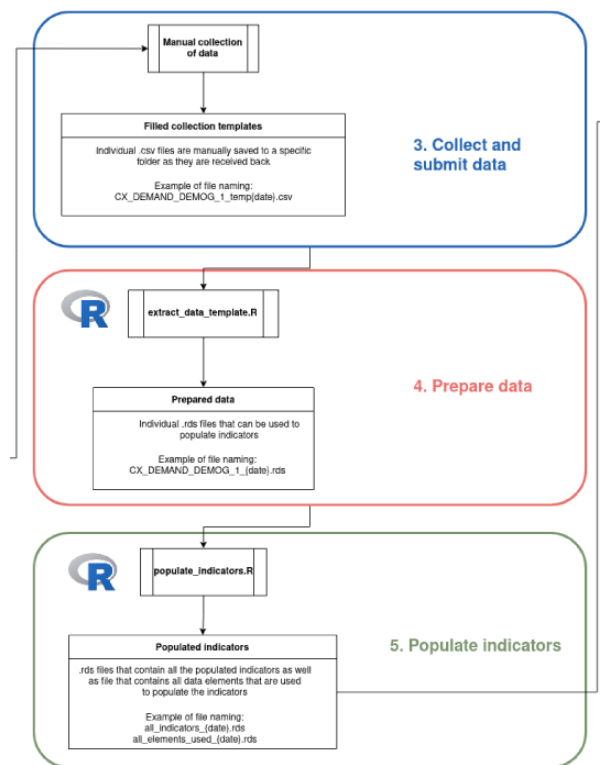


Steps 1-2: Metadata and collection templates



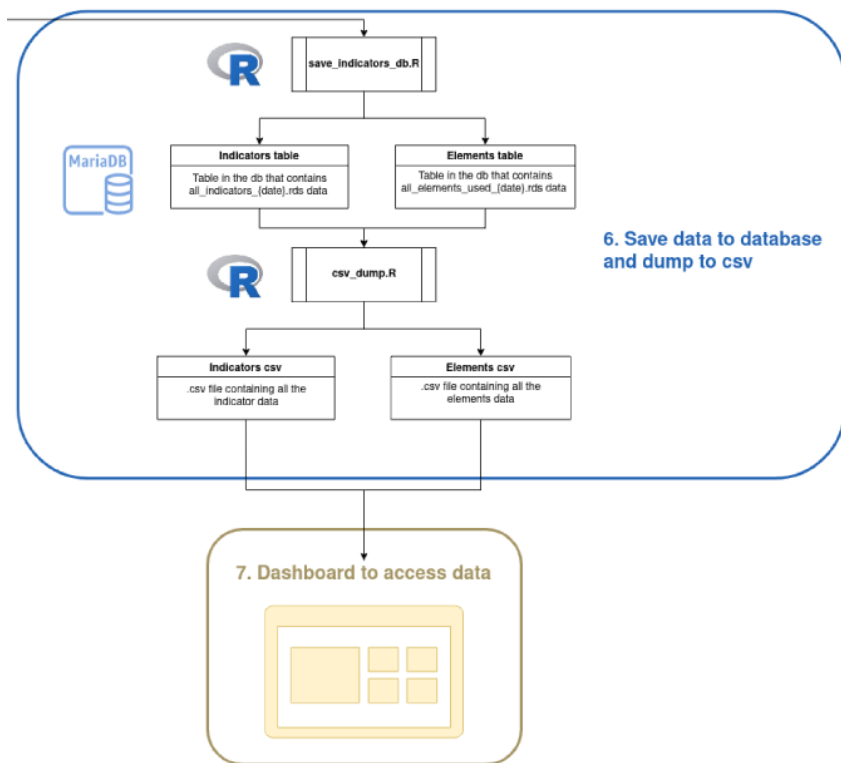
- each indicator has its own individual `.csv` file
 - assists with version control & allows multiple people to work on different indicators' metadata simultaneously
- there is a single `.csv` file containing all the data elements
 - ensures there is no unnecessary duplication of data elements
- format of the indicator and metadata elements `.csvs` must be kept constant
- `make_template.R` creates one `.csv` collection template per indicator (easy assignment of collection)

Steps 3-5: Collect, submit, prepare, and populate



- collection templates are filled in manually and saved in a specific location on the server
 - saved with a **specific** naming convention
 - converted to `.rds` format to facilitate downstream processing using the `extract_data_template.R` script
- the `populated_indicators.R` script contains a function that reads in the converted data and populates all of the indicators
 - this scripts transforms the metadata from sentences into `code`

Steps 6-7: Save and disseminate



- prepared and populated data is saved in a database
 - ensures everyone is always looking at the *same* version of the data
 - database is kept up to date by means of a script scheduler (`crontab`) that runs through the pipeline on a regular basis
- data from database is written to `.csv` files that are used to populate the dashboard
 - dashboard is kept up to date by means of a script scheduler
 - this setup affords development of an API further down the line

Why do we need metadata?

- Transparency and reproducibility

Everyone knows what the indicator is measuring and should be able to produce the same value when using the same dataset. This reduces uncertainty during both the data collection and data dissemination stages.

- Automation

Unfortunately, we are error-prone and populating indicators by hand in Excel will lead to mistakes. Once developed, metadata can be translated into code to minimize the occurrence of such errors.

Automation makes it easier to onboard new members to the team, assisting with the longevity of the project

- Identification of data needs

Metadata documents and formalizes data collection requirements and can inform engagement strategies that aim to realize efficiencies during data collection (templates, APIs, etc.)

Developing metadata

Start with a framework

The hardest part is to start from scratch - no indicators, no elements, just an idea of what data you possess / what is out there.

A metadata framework can help to guide you to what indicators you need - it assists with identifying indicators of interest (i.e. coming up with indicator names). Once you have a name, you can define it.

A good framework can also show you which indicators are unnecessary

- This project will make use of the MSI metadata framework developed by CAHF, Reall, and 71point4
 - housing focus makes it ideal for this project
- All indicators must be developed within this framework
 - this means that the indicator metadata developed by the KNBS must contain the same fields as that contained in the MSI metadata

The MSI framework

The **MSI framework** classifies 115 indicators into three tiers - categories, sub-categories, and components

- Each category has specific sub-categories
- Each sub-category has specific components

There are two categories: Value Chain (89 indicators) and Context (26 indicators)

- Value Chain sub-categories: Land and Infrastructure, Construction and Investment, Sales and Rental, Maintenance and Management
 - 16 components
- Context sub-categories: Enabling Environment, Economic Environment, Demand
 - 3 components

MSI indicators as per the three-tier framework

Category	Sub-category	Component	Number of indicators
Value chain indicators	Land & infrastructure	Land assembly	8
		Land title	7
		Infrastructure	8
	Construction & investment	Stock	6
		Flow	3
		Industry	5
		Building materials	2
		Process	2
	Sales & rental	Ownership	7
		Rental	2
		Transactions	8
		Finance	20
		Affordability	5
	Maintenance & Management	Home improvements	2
		Municipal management	2
		Finance	2
Context indicators	Enabling environment	Operating environment	6
	Economic environment	Macroeconomic indicators	13
	Demand	Demographics	7

MSI indicators as per the three-tier framework

- The three tiers help to focus efforts
- This makes it easier to develop indicators of interest (not a blank slate anymore)
- Also easier to identify gaps
 - Where are we thin on indicators?

1.2 Land title	Total number of residential properties with a title deed
1.2 Land title	Number of procedures to register residential property
1.2 Land title	Name of residential property registration procedure that takes the longest to complete
1.2 Land title	Time to register residential property (days)
1.2 Land title	Cost to register residential property
1.2 Land title	World Bank DBI transparency of information index ranking: Africa
1.2 Land title	World Bank DBI transparency of information index ranking: Global
1.3 Infrastructure	% of residential development projects where developers are paying for bulk infrastructure or the building of roads
1.3 Infrastructure	% of households without access to improved drinking water services
1.3 Infrastructure	% of households without access to improved sanitation services
1.3 Infrastructure	% of households without access to electricity
1.3 Infrastructure	% of households living in dwellings built using durable building materials (walls and roof) with inadequate services

What makes an indicator?

1. It is derived from one or more data elements

Every indicator is made up of at least one data element. If it is made up of one data element, then the indicator and the element will have the same value.

Indicator	Number of data elements
Number of building works completed	??

Year	KSh Million	
	Building Plans Approved	Building Works Completed ¹
2000....		
2001....		
2002....		
2003....		
2004....		
2005....		
2006....		
2007....		
2008....		
2009....		
2010....		
2011....	112,842.80	42,464.10
2012....	135,128.20	48,273.70
2013....	190,646.50	52,276.00
2014....	205,423.90	59,519.70
2015....	215,211.00	70,867.40
2016....	308,361.40	77,749.70
2017....	240,752.00	86,128.40
2018....	210,296.71	90,127.40
2019....	207,624.90	93,982.30
2020*...	153,575.40	100,041.30

Source: Nairobi City County

*Provisional

¹ Exclude extensions

What makes an indicator?

1. It is derived from one or more data elements

Every indicator is made up of at least one data element. If it is made up of one data element, then the indicator and the element will have the same value.

Indicator	Number of data elements
Number of building works completed	1

Year	KSh Million	
	Building Plans Approved	Building Works Completed ¹
2000....		
2001....		
2002....		
2003....		
2004....		
2005....		
2006....		
2007....		
2008....		
2009....		
2010....		
2011....	112,842.80	42,464.10
2012....	135,128.20	48,273.70
2013....	190,646.50	52,276.00
2014....	205,423.90	59,519.70
2015....	215,211.00	70,867.40
2016....	308,361.40	77,749.70
2017....	240,752.00	86,128.40
2018....	210,296.71	90,127.40
2019....	207,624.90	93,982.30
2020*...	153,575.40	100,041.30

Source: Nairobi City County

*Provisional

¹ Exclude extensions

What makes an indicator?

1. It is derived from one or more data elements

Table 11.10: Quarterly Building Cost Index, 2020
Base Period Dec 2019=100

No	Product	Weight	Dec-21
1	Cement & Lime	14.18	96.32
2	Hard core	1.20	103.64
3	Quarry products (waste, dust & murrum)	1.72	103.90
4	Sand	5.14	93.93
5	Ballast	4.50	97.96
6	BRC Mesh & Steel Reinforcement Bars	10.33	131.81
7	Stones (Machine Cut & Foundation Stones)	4.42	95.35
8	Damp Proofing & Anti-termite	0.12	91.84
9	Timber & Wood	3.70	99.57
10	Paving blocks	0.14	120.61
11	Roofing materials (Iron sheets, Tiles, Gutters, down-pipe & Nails)	3.69	101.61
12	Doors	0.54	100.68
13	Windows	0.26	103.42
14	Glass & glass putty	0.41	114.18
15	Locks & iron mongery	0.24	105.31
16	Tiles (Wall & Floor)	1.25	104.53
17	Chip boards & MDF	0.78	98.66
18	Paints	1.61	102.13
19	Sanitary fittings	0.16	103.18
20	Water fittings	0.35	101.50
21	Water wastes	0.12	102.91
22	Electrical fittings	0.51	102.79
Total materials		55.36	104.55

23	Equipment-Concrete Mixer	5.16	97.99
24	Equipment-Concrete poker / Vibrator	2.36	103.71
25	Equipment-Excavator & Pedestrian Roller	2.64	100.01
26	Compressors	3.04	99.45
Total equipments		13.21	99.76
27	Transport	5.99	108.90
28	Fuel	3.99	102.30
Total fuel & transport		9.98	106.26
29	Casual	6.67	106.99
30	Watchman	2.01	106.49
31	Plumber/Electrician	0.52	102.38
32	Machine /plant operators	0.91	100.61
33	Carpenter/Painter/Welder/Mechanic	3.53	104.56
34	Mason/foreman	7.80	104.43
Total labour		21.45	105.23
Overall Building Cost Index		100.00	104.23

NOTE:-Prices are collected a quarterly basis on every 15th of the mid-month of the quarter

Indicator

Number of data elements

Quarterly building cost index ??

What makes an indicator?

1. It is derived from one or more data elements

Table 11.10: Quarterly Building Cost Index, 2020
Base Period Dec 2019=100

No	Product	Weight	Dec-21
1	Cement & Lime	14.18	96.32
2	Hard core	1.20	103.64
3	Quarry products (waste, dust & murrum)	1.72	103.90
4	Sand	5.14	93.93
5	Ballast	4.50	97.96
6	BRC Mesh & Steel Reinforcement Bars	10.33	131.81
7	Stones (Machine Cut & Foundation Stones)	4.42	95.35
8	Damp Proofing & Anti-termite	0.12	91.84
9	Timber & Wood	3.70	99.57
10	Paving blocks	0.14	120.61
11	Roofing materials (Iron sheets, Tiles, Gutters, down-pipe & Nails)	3.69	101.61
12	Doors	0.54	100.68
13	Windows	0.26	103.42
14	Glass & glass putty	0.41	114.18
15	Locks & iron mongery	0.24	105.31
16	Tiles (Wall & Floor)	1.25	104.53
17	Chip boards & MDF	0.78	98.66
18	Paints	1.61	102.13
19	Sanitary fittings	0.16	103.18
20	Water fittings	0.35	101.50
21	Water wastes	0.12	102.91
22	Electrical fittings	0.51	102.79
	Total materials	55.36	104.55

23	Equipment-Concrete Mixer	5.16	97.99
24	Equipment-Concrete poker / Vibrator	2.36	103.71
25	Equipment-Excavator & Pedestrian Roller	2.64	100.01
26	Compressors	3.04	99.45
	Total equipments	13.21	99.76
27	Transport	5.99	108.90
28	Fuel	3.99	102.30
	Total fuel & transport	9.98	106.26
29	Casual	6.67	106.99
30	Watchman	2.01	106.49
31	Plumber/Electrician	0.52	102.38
32	Machine /plant operators	0.91	100.61
33	Carpenter/Painter/Welder/Mechanic	3.53	104.56
34	Mason/foreman	7.80	104.43
	Total labour	21.45	105.23
	Overall Building Cost Index	100.00	104.23

NOTE: Prices are collected a quarterly basis on every 15th of the mid-month of the quarter

Indicator

Number of data elements

Quarterly building cost index 1 / 8 / 64

What makes an indicator?

1. It is derived from one or more data elements

Data elements differ from indicators in that we collect data elements, but we populate indicators - the elements are the micro and the indicators the macro, the indicator is the objective, its elements are the means

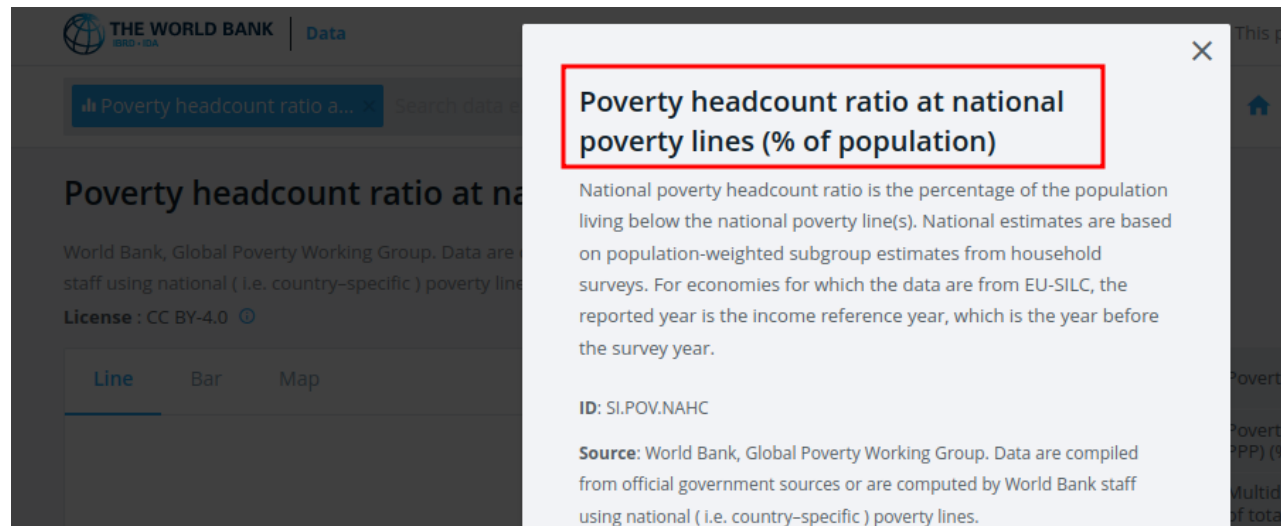
- Because indicators are a composite of data elements, we require much of the same metadata for data elements as we do for indicators
- Therefore, when developing metadata, you need to simultaneously develop the indicator metadata as well as the data element metadata
 - We advise on having a single `data_elements_metadata.csv` field for the data elements metadata
 - One `.csv` file per indicator for the indicator metadata
 - This allows one person to 'own' the development of an indicator's metadata without running into version control issues

What makes an indicator?

2. It has a name

The indicator name should be reasonably short, but with enough detail that users know what it is trying to measure.

We denote indicator names with the `indicator_name` field and data element names with the `data_element_name` field



The screenshot shows the World Bank Data portal interface. The main heading is 'Poverty headcount ratio at national poverty lines (% of population)'. Below the heading, there is a description: 'National poverty headcount ratio is the percentage of the population living below the national poverty line(s). National estimates are based on population-weighted subgroup estimates from household surveys. For economies for which the data are from EU-SILC, the reported year is the income reference year, which is the year before the survey year.' The source is listed as 'World Bank, Global Poverty Working Group. Data are compiled from official government sources or are computed by World Bank staff using national (i.e. country-specific) poverty lines.' The ID is 'SI.POV.NAHC'. The license is 'CC BY-4.0'. The chart type is set to 'Line'.

What makes an indicator?

3. It has a key

The indicator key serves as the unique identifier for each indicator. This field is useful for automation and ensuring everyone is talking about the same indicator - sentences can get confusing

We denote indicator keys with the `indi_key` field and data element keys with the `de_key` field

The screenshot shows the World Bank Data portal interface. The main heading is 'Poverty headcount ratio at national poverty lines (% of population)'. Below the heading, there is a description: 'National poverty headcount ratio is the percentage of the population living below the national poverty line(s). National estimates are based on population-weighted subgroup estimates from household surveys. For economies for which the data are from EU-SILC, the reported year is the income reference year, which is the year before the survey year.' The indicator ID 'SI.POV.NAHC' is highlighted in a red box. Below the ID, the source is listed as 'World Bank, Global Poverty Working Group. Data are compiled from official government sources or are computed by World Bank staff using national (i.e. country-specific) poverty lines.' The license is 'CC BY-4.0'.

What makes an indicator?

4. It has a definition

As far as is possible, the indicator definition should dispel any uncertainty around what the indicator measures and what it doesn't measure. Be as explicit as you can here.

% of households without access to improved sanitation services

Metadata field	Value
Indicator key	VC.LAND.INFRA.3
Indicator name	% of households without access to improved sanitation services
Definition	The share of households without access to an improved sanitation facility. According to DHS 7, these include: flush - to piped sewer system; flush - to septic tank; flush - to pit latrine; flush - don't know where; pit latrine - ventilated improved pit (VIP); pit latrine - with slab; composting toilet
Formula	$\frac{\text{Number of households without access to improved sanitation services}}{\text{Number of households}}$
Notes	
Collection frequency	Annual
Aggregations of interest	National, Urban, Main urban centre, B40, Developer
Suggested sources	Statistics Bureau/DHS

What makes an indicator?

5. A formula *may* be required to derive it

Indicators composed of multiple elements will require a formula to be derived. This metadata field captures that formula and is critical for accurate translation of the metadata to `code`.

Use the keys of the *data elements* to contextualize the formula. This will make it easier down the line to translate the formula into `code`

indi_key	indicator_name	definition	de_key	formula
KNBS_CONS_9	Quarterly building cost index	The Residential and Non Residential building cost index as reported by CIPI. Base period Dec 2019 = 100.	CON_9; CON_10, CON_11; CON_12; CON_13; CON_14; CON_15; CON_16	$(CON_9 \times CON_10) + (CON_11 \times CON_12) + (CON_13 \times CON_14) + (CON_15 \times CON_16)$

What makes an indicator?

5. A formula *may* be required to derive it

The `data_elements_metadata.csv` can be used to understand the formula in more detail if required

de_key	data_element_name	definition	formula
CON_9	Building cost total materials index	The Residential and Non Residential building cost index value for the Total materials component as reported by CIPI. Base period Dec 2019 = 100.	
CON_10	Building cost total materials weight	The Residential and Non Residential building cost weight value for the Total materials component as reported by CIPI. Base period Dec 2019 = 100.	
CON_11	Building cost total equipments index	The Residential and Non Residential building cost index value for the Total equipments component as reported by CIPI. Base period Dec 2019 = 100.	
CON_12	Building cost total equipments weight	The Residential and Non Residential building cost weight value for the Total equipments component as reported by CIPI. Base period Dec 2019 = 100.	

Additional metadata to be developed by the KNBS

The following additional metadata fields must also be developed by the KNBS. **NOTE** that they are for the indicator metadata only.

- aggregation
 - This field indicates the level of aggregation at which the indicator is to be calculated: National, Urban, Main urban centre, B40
 - Multiple values are possible and should be separated using the ; delimiter
- notes
 - Additional notes to assist with understanding the context of the indicator, how the data it requires can be collected, how to overcome difficulties with data availability, etc – **not needed for all indicators**
- category, sub-category, component
 - MSI framework three-tier classification levels that the indicator relates to

Indicator metadata template

Based on the above, the indicator metadata template will have the following fields

Metadata field	Meaning
indi_key	Unique identifier of the indicator
indicator_name	Short name of the indicator - should provide good indication of what it is measuring
definition	Detailed definition of the indicator - the more the better!
de_key	The unique identifiers of the data elements used to populate the indicator
formula	Explains how the data elements that make up an indicator are combined to derive the indicator – not completed for all indicators
aggregation	Aggregation at which the indicator is calculated: National, Urban, Main urban centre, B40
notes	Notes to assist with understanding the context of the indicator, how the data it requires can be collected, how to overcome difficulties with data availability, etc – not needed for all indicators
category	MSI framework top-tier classification that the indicator relates to
sub_category	MSI framework second-tier classification that the indicator relates to
component	MSI framework third-tier classification that the indicator relates to

Indicator metadata template

An example metadata file for the indicator associated with `indi_key = KNBS_CONS_9` is provided below

- The naming convention for the indicator metadata files are: `<indi_key>.csv`
- For this example the file name would be `KNBS_CONS_9.csv`

indi_key	indicator_name	definition	de_key	formula	aggregation	notes	category	sub_category	component
KNBS_CONS_9	Quarterly building cost index	The Residential and Non Residential building cost index as reported by CIPI. Base period Dec 2019 = 100.	CON_9; CON_10, CON_11; CON_12; CON_13; CON_14; CON_15; CON_16	$(CON_9 \times CON_10) + (CON_11 \times CON_12) + (CON_13 \times CON_14) + (CON_15 \times CON_16)$	National		Value Chain	Construction and investment	Building materials

Data element template

The data element metadata template has fewer fields than the indicator metadata template - an example is provided below

- There is just one data element metadata file containing all of the data elements
- This file should be named `data_element_metadata.csv`

de_key	data_element_name	definition	formula
CON_9	Building cost total materials index	The Residential and Non Residential building cost index value for the Total materials component as reported by CIPI. Base period Dec 2019 = 100.	
CON_10	Building cost total materials weight	The Residential and Non Residential building cost weight value for the Total materials component as reported by CIPI. Base period Dec 2019 = 100.	
CON_11	Building cost total equipments index	The Residential and Non Residential building cost index value for the Total equipments component as reported by CIPI. Base period Dec 2019 = 100.	
CON_12	Building cost total equipments weight	The Residential and Non Residential building cost weight value for the Total equipments component as reported by CIPI. Base period Dec 2019 = 100.	

Metadata checklist

- Create the metadata template for indicators and elements
- Fill in template
 - What is the name of the indicator?
 - Where does it fit in the framework?
 - assign tier 1, 2, 3
 - What elements are required to generate it?
 - Are any of them already in the elements sheet? - if not, then develop it's metadata first in `data_element_metadata.csv`
 - Are multiple elements required? If yes then fill in the formula using the `de_keys` of its data elements
 - What aggregations are available?
- Review metadata for errors



What happens next?



Data collection template

Once the metadata is complete, we can start collecting the data.

The data collection templates can be generated automatically by leveraging the standardized metadata formats, i.e. we can write a script to create the data collection templates.

- These will be output as `.csv` files
- There will be one `.csv` file per indicator

Having one `.csv` file per indicator improves the efficiency of the data collection efforts

We will provide you with this script and will explain it in more detail during the second workshop.

- It is not necessary to write or change this script if the metadata formats are adhered to

Automated indicator population example

- You need to know some basic `R` and `tidyverse`
 - We will be writing functions to automatically populate the indicators
- First we check that everything we need is available

```
## CX.DEMAND.DEMOG.1 -----  
if (indi_key == "CX.DEMAND.DEMOG.1") {  
  check_all_elems <- unique(df$de_key)  
  
  if(any(is.na(df$value))) {  
    indi_df <- "Missing values"  
  }  
  
  if(check_all_elems != "DE.1" || length(check_all_elems) != 1 || length(df$de_key) > 1){  
    indi_df <- "Incorrect data element or multiple data elements provided"  
  }  
}
```

Automated indicator population example

Then we populate the indicators

```
if(check_all_elems == "DE.1" && length(df$de_key) == 1 && !any(is.na(df$value))){  
  indi_df <- tibble(  
    indi_key = indi_key,  
    aggregation = aggregation,  
    year = df$year,  
    value = df$value,  
    source_dataset = df$source_dataset,  
    collection_note = df$collection_note,  
    url = df$url  
  )  
  na_cols <- names(which(sapply(indi_df, function(x) any(is.na(x)))))  
  warning(glue("The following columns have NA values - check your input data frame: {na_cols}"))  
}  
  
return(indi_df)
```

What do you need?

- RStudio on your computer - [installation instructions](#)
 - R and RStudio are open source software for statistical programming
 - We will be writing scripts together to automate the population of the indicators
- Complete [this free tidyverse course](#)
 - Tidyverse is a collection of R software packages
 - These packages encapsulate a programming philosophy that is reasonably easy to pick-up
 - We will be using these packages to develop the scripts needed to populate the indicators
- Complete chapters 17 to 21 of of R for Data Science - [available here](#)
 - This book is a great introduction to using R for Data Science
 - Chapters 17 - 21 introduce users to writing functions - we will be using these to automate the data pipeline