

Methodology update

Now that we've moved into micro-level data gathering involving country representatives and other stakeholders, we need to set some new systems in place. In addition, we should start thinking more explicitly about outputs of the crosswalk—i.e. what is the end product we'd like to produce? In what follows we discuss our current data developments as well as suggestions for possible outputs.

Data gathering

There are two main microdata types we'll be working with—project data and institutional data. Project data has to do with development projects and programs. Ideally we'd like to gather information about project characteristics, inputs, outputs, and outcomes as well as any monitoring and evaluation (M&E) reports. This data is easier to find since many donors organizations are promoting open data procedures and posting more and more data online.

Institutional data is basic reporting information about what's happening on the ground. Different institutions such as health clinics or farmer's unions are ideally already gathering this type of information and reporting it to local government agencies. This also includes surveys such as national or local censuses. This data will be harder to find since it comes from the field where capacity for organizing and storing data is likely not incredibly developed. Thus, most of what is publically available is already aggregated at national levels.

We have limited our data gathering to the case countries: Ghana, Tanzania, and Sri Lanka (in health and agriculture). Within these, we are gathering both data types from a variety of sources. For the project/ evaluation-type data, we've compiled datasets and other project information from the World Bank, MCC, and others. This is depicted in the following table.

Table 1: Project data
[Forthcoming...]

For institutional data, we've gathered some census information, but are still exploring the data-gathering systems for each country and will then make decisions about who to contact and request data from. Based on our conversations in the consultation meeting, local government officials may be our best bet for micro-level information. Microdata is important since it can be flexible in terms of users—i.e. local actors can use it for situational decision-making and reporting and it can also be aggregated for decisions up the chain. The following is the institutional data we have so far.

Table 2: Institutional data
[Forthcoming...]

Data organization

At its core, the crosswalk is a data organization activity and so our main output will either be a data organization system itself or at least a couple of solid recommendations for possible systems. So far we've been operating with the World Development Indicator

method—i.e. a system of indicators themselves grouped by locations (countries) and years. The following figure shows this.

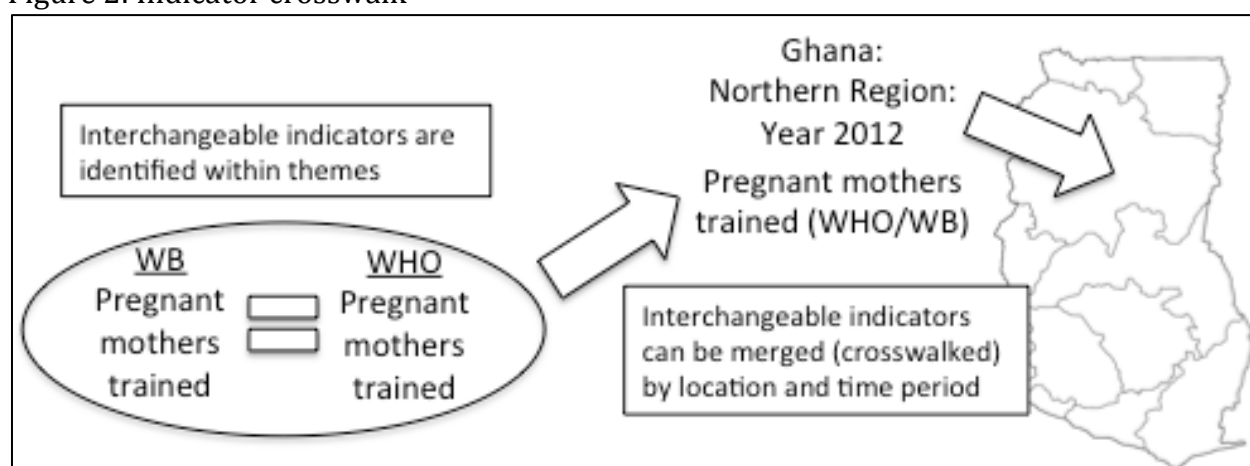
Figure 1: The World Development Indicators

The screenshot shows the WDI interface with the following details:

- Database:** Available 1 | Selected 1
- Country:** Available 249 | Selected 2
- Series:** Available 1343 | Selected 0
- Search:** Enter Keywords for Search
- Alphabetical Index:** 2 A B C D E F G H I L M N O P Q R S T U V W
- Series List:**
 - ☐ 2005 PPP conversion factor, GDP (LCU per international \$)
 - ☐ 2005 PPP conversion factor, private consumption (LCU per international \$)
 - ☐ Access to electricity (% of population)
 - ☐ Access to electricity, rural (% of rural population)
 - ☐ Access to electricity, urban (% of urban population)
 - ☐ Access to non-solid fuel (% of population)
 - ☐ Access to non-solid fuel, rural (% of rural population)

Our contribution to this system is that we would include data from multiple organizations and also location information at subnational levels. Some of these indicators would overlap, but we would link them as part of the crosswalk. Thus, you could select from a list of indicators and view or aggregate them by location and year.

Figure 2: Indicator crosswalk



However, this system has limitations, the foremost being that it only crosswalks data based on indicator names and metadata—it does not group together indicators based on user

interest or causal relationships. This means that indicators can be organized by keywords, themes, and locations, but there is no differentiating between inputs, outputs, outcomes, and evaluations. It is basically just a data assembly method.

However, there are horizontal and vertical relationships in results data that we may want to preserve. Vertically, data originates in projects and institutions and then is sent up the chain through evaluations and summary reports. Consider, for example, the World Bank's calculation of a single indicator: Access to electricity.

Table 3: World Development Indicators metadata

Indicator Name	Long definition	Source	Aggregation method	Statistical concept and methodology
Access to electricity (% of population)	Access to electricity is the percentage of population with access to electricity. Electrification data are collected from industry, national surveys and international sources.	World Bank, Sustainable Energy for all (SE4ALL) database from World Bank, Global Electrification database.	Weighted average (annual)	Data for access to electricity are collected among different sources: mostly data from nationally representative household surveys (including national censuses) were used. Survey sources include Demographic and Health Surveys (DHS) and Living Standards Measurement Surveys (LSMS), Multi-Indicator Cluster Surveys (MICS), the World Health Survey (WHS), other nationally developed and implemented surveys, and various government agencies (for example, ministries of energy and utilities). Given the low frequency and the regional distribution of some surveys, a number of countries have gaps in available data. To develop the historical evolution and starting point of electrification rates, a simple modeling approach was adopted to fill in the missing data points - around 1990, around 2000, and around 2010. Therefore, a country can have a continuum of zero to three data points. There are 42 countries with zero data point and the weighted regional average was used as an estimate for electrification in each of the data periods. 170 countries have between one and three data points and missing data are estimated by using a model with region, country, and time variables. The model keeps the original observation if data is available for any of the time periods. This modeling approach allowed the estimation of electrification rates for 212 countries over these three time periods (Indicated as "Estimate"). Notation "Assumption" refers to the assumption of universal access in countries classified as developed by the United Nations.


This indicator is calculated using, "Electrification data ... collected from industry, national surveys and international sources"—i.e. micro-level indicators from a variety of sources.


Horizontally, data (from projects especially) can be separated into input indicators that produce outputs, which through evaluation we can ideally tie to outcomes. Teacher training, for example, is an input that produces a certain number of trained teachers. An evaluation of this teacher-training program may also note that it contributed to improved education in the area.

There are various ways we could capture these kinds of relationships. The most common method among development organizations is to build databases at the project level. Each project has its own page with evaluation and input information as well as a dataset of outputs. For example, a page for a Rwandan program in the MCC database is shown in the following figure.

Figure 3: MCC's project database

Rwanda - Threshold Impact



Reference ID	DDC-MCC-RWA-THRESHOLD-MPR-2014-v1.1
Year	2011
Country	Rwanda
Producer(s)	Mathematica Policy Research
Sponsor(s)	Millennium Challenge Corporation - MCC -
Metadata	 Documentation in PDF





Created on	Oct 15, 2014
Last modified	Sep 11, 2015
Page views	7789
Downloads	654

[Documentation](#)
[Study Description](#)
[Data Description](#)
[Get Microdata](#)





Documentation

Download the questionnaires, technical documents and reports that describe the survey process and the key results for this study.

Questionnaires

 Baseline Questionnaire	 155.16 KB
 Follow-up Questionnaire	 545.52 KB

Reports

 Baseline Report	 679.72 KB
 Evaluation Design Report	 259.87 KB

Note that project documentation and evaluations are included as well as a “Get Microdata” link which allows a user to download the project’s data. This type of system would also be an option, our innovation again being that we would compile information from a variety of donors. However, this too has already been done to some extent. Note the following figure showing the World Bank’s Microdata Library.


Figure 4: World Bank Microdata Catalog



World Bank Country Survey 2012
Afghanistan, 2012

By: Public Opinion Research Group - The World Bank Group
Collection: **The World Bank Group Country Opinion Survey Program (COS)**

Created on: Mar 14, 2014 Last modified: Mar 14, 2014



Global Financial Inclusion (Global Findex) Database 2011
Afghanistan, 2011

By: Development Research Group, Finance and Private Sector Development Unit - World Bank
Collection: **Global Financial Inclusion (Global Findex) Database**

Created on: Dec 12, 2012 Last modified: Apr 15, 2015 Citations: 3



Multiple Indicator Cluster Survey 2010-2011
Afghanistan, 2010-2011

By: Central Statistics Organization - Government of the Islamic Republic of Afghanistan, United Nations Children's Fund
Collection: **UNICEF Multiple Indicator Cluster Surveys (MICS)**

Created on: Jan 06, 2014 Last modified: Jan 06, 2014



Mortality Survey 2010
Afghanistan, 2010

By: Indian Institute for Health Management Research (IIHMR), Central Statistics Organization (CSO)
Collection: **MEASURE DHS: Demographic and Health Surveys**

Created on: Feb 26, 2013 Last modified: Sep 26, 2013

Note that each program has a “Collection” link which notes where the dataset originated. This database includes microdata from a variety of organizations including UNICEF, DHS, etc. However, this allows for even less data grouping since each dataset is tied to an individual project. Thus, clustering similar data would be an arduous task and aggregating it even more so.

Our proposal

For these reasons we propose a hybrid method that looks more like the WDIs, but includes variable tags based on horizontal and vertical information. This requires more human intervention, but allows us to preserve relationships in the data. For example, the teacher-training program described above would now have an “improved local education” tag, which could be grouped with other indicators that also improved local education. This would allow users to search for indicators based on their own outcomes of interest instead of just similar names or themes.

Crosswalking would still occur at the indicator level, but can now be combined with a common language of outcomes among donors. Similarly, vertical information can also be recorded. For example, local income or CPI indicators could have a “GDP” tag, meaning that these are among the variables used in the composite GDP statistic. This would allow users to also search for micro-level indicators in order to drill down into macro-level variables.

Moving into this kind of system requires a lot more qualitative work on our part in two main ways. First, we need to come up with systematic ways to search through evaluation information and quantify it. This is mostly a process for project-type data—different donors have different priorities when it comes to evaluation reports. DHS, for example, doesn’t even have its own set of evaluation criteria, but instead formulates them based on the priorities of the project itself. Thus, we will need to come up with a consistent way in which to crosswalk evaluation information across donor agencies.

Second, we have to research the metadata of macro-level indicators to ascertain their micro-level inputs. This is more common among institutional data types—higher levels of government may have different standards for core indicators and may compile them in different ways. Hopefully by tracking these indicators at the micro-level, we can come up with a “macro-indicator translation” system—e.g. a way to calculate Ghana’s GDP using Tanzania’s method and vice versa.

The deliverable

Since we plan to organize the data as described above, we think the ideal deliverable would be a searchable online database like the WDIs. (It’s kind of like a [Metacritic](#), but for data). However, since we propose to include both horizontal and vertical relationships in the data, it would be important to talk in more detail about how to present the information in an accessible way. As we begin searching through project and evaluation documentation, this should give us a better indication of how to organize the information using links, drop-down menus, etc.