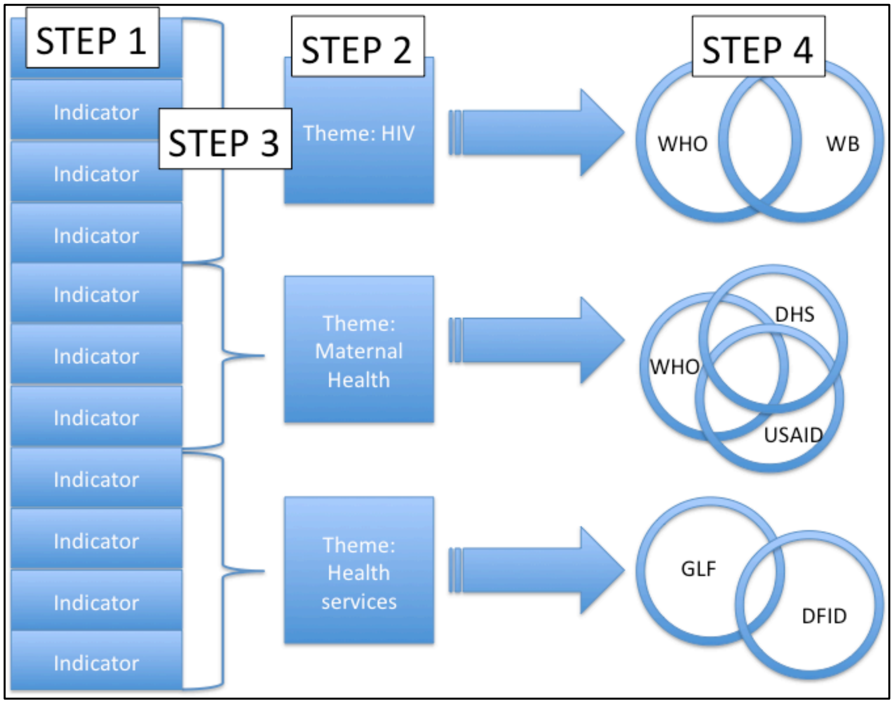
Development Gateway
Results Data Crosswalk
Executive Summary (2 pages)

The main goal of this project is to compare results information across organizations, countries, years, and other relevant criteria. This should allow for better data organization and use as well as improved coordination among development associations. There are four main steps to this exercise as shown in the figure below:

Figure 1: Results data methodological approach



In Step 1, we gather together indicators and results from a variety of organizations. Although these indicators may be measuring the same thing, we'll need to perform some intermediary analyses to find out if this is the case. For this reason, in Step 2, we create indicator categories—"themes"—which will cluster together indicators of similar types (e.g. HIV, irrigation, etc.). In Step 3, we use Machine Learning algorithms to pull our raw list of indicators into the appropriate groups. This makes our overall indicator list (numbering already in the thousands with only a sub-set of development partners) more manageable. Then in Step 4 we can look in each theme for specific indicators that overlap. These overlapped indicators will be documented, and we will subsequently assess what calculations or corrections need to be made such that they are directly comparable (e.g. translating actual increases into % increases).

Progress to-date

Step 1: Gathering data

So far we have been able to compile a fairly large list of indicators (over 4,000) from a variety of health organizations. Each of these has its own unique measures of health, but even a cursory review of the raw data list shows that there is considerable overlap, particularly in the areas of maternal health and HIV.

Step 1: Moving forward

We need lists of indicators and, ideally, accompanying metadata from both health and agriculture organizations. Most of this is available online for larger organizations, but we will need advice on which smaller organizations may have useful information and where their data can be accessed. We also need a reliable way to extract results data, which is mostly contained in PDF reports. In the short term, the team is hiring 1-2 interns to manually extract this information.

Step 2: Creating themes and classification schema

We have experimented with a few theme-creation methods, but have ultimately settled on an iterative approach. In health, for example, we use the Top 100 (a list compiled by the WHO and others of the most important health variables to gather in the field) as a baseline, verifying its accuracy and usefulness using word frequency analysis. Then, we can update these results based on our familiarity with the data and using advice from experts in the field. By combining the strengths of each categorization type, we can ensure the formation of detailed groups. This will be important for the next phase of our analysis—classifying new indicators based on our list of themes.

Step 2: Moving forward

We need a solid list of key terms and themes. To get this we will seek feedback on the lists we are currently using and what they might be missing or overemphasizing. This will be especially important as we move into agriculture since there is much less consensus on high-priority indicators and key themes.

Step 3: Data classification

Once we have a working list of key themes and indicators, we can start classifying new indicator data. For this we will be using various Machine Learning methods such as Support Vector Machines (SVM) and Neural Networks. The particular specifications of each modeling type are beyond the scope of this methodology brief; suffice to say they are methods that "learn" from one set of classifications and then apply these learned rules to new data. This is why a solid list of key indicators is important—the better and more detailed our initial list, the better our algorithms will be at predicting the categories of new indicator lists. So far we have tried a host of different models and are continually working towards greater accuracy.

Step 3: Moving forward

Presently, the team is focused on improving our classification models. Feedback on the first two steps—particularly with compiling a good list of themes and keywords—will be invaluable to this process.

Step 4: Data crosswalking

The final step is to directly compare most-similar indicators across organizations. To do this, we have been using "edit distance" algorithms, which compare indicator names themselves. A quick example of this can be seen in the table below. Once we have a list of similar indicators, we can standardize them across organizations using information from metadata.

Table 1: Similar indicators based on name

Indicator #1	**Org. #1**	**Indicator #2**	**Org. #2**	**Distance**
TB/HIV mortality rate	Global Fund	HIV mortality rate	WHO	3
HIV prevalence among men	DHS	HIV prevalence	WHO	10

Step 4: Moving forward

We need to make a working example of this type of data crosswalk using a small sample of similar indicators. We can then make this into a workflow and involve one or more coders—they will be in charge of identifying similar indicators and documenting how they can be compared across organizations.

Full analysis (9 pages)

This section goes more in-depth into the data organization, classification, and analysis process. Preliminary results at each stage are discussed.

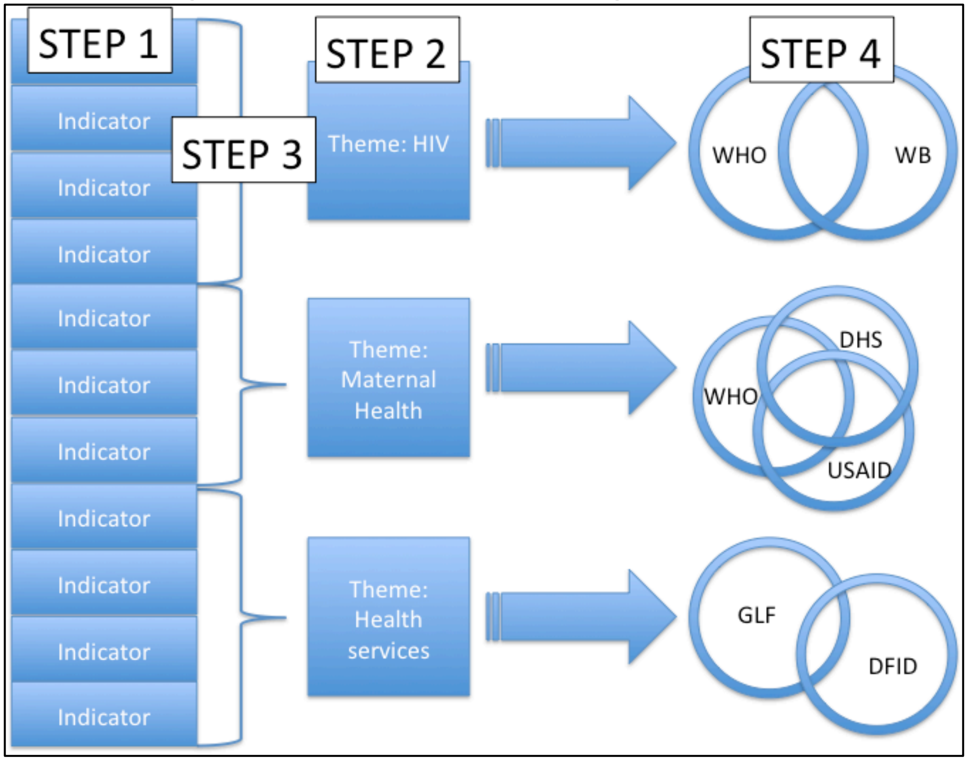
Definitions

- **"Development organizations"—**organizations involved in international development, particularly those who fund health and/ or agricultural initiatives. Examples include the WHO, World Bank, and USAID.
- **"Results information"—**information reported to these organizations as to the results of their initiatives. For our purposes, results data spans both "outputs" and "outcomes." Output data reports on service delivery – number of teachers trained, number of children vaccinated, number of seeds distributed, etc. Outcome data tracks longer-term results, like changes in illness rates, crop yields, and so on. These data are collected in a number of formats, including project monitoring or progress reports, surveys, and formal evaluations. For example, a USAID implementing partner may submit a monthly progress report on its HIV prevention activities to a USAID mission. This report may include a list of numerical output indicators and a narrative description of project performance. We focus on aggregated numerical indicator data for this crosswalk activity.
- **"Comparability"—**the degree to which results information from similar initiatives can be compared. This could be across or within organizations—for example, the World Bank may have HIV prevention projects in both Burundi and Afghanistan while WHO has similar projects in Thailand and Nigeria. Comparability is the degree to which reported information from all of these projects can be grouped together.

General approach

There are four main steps to this exercise as shown in the figure below:

Figure 1: Results data methodological approach



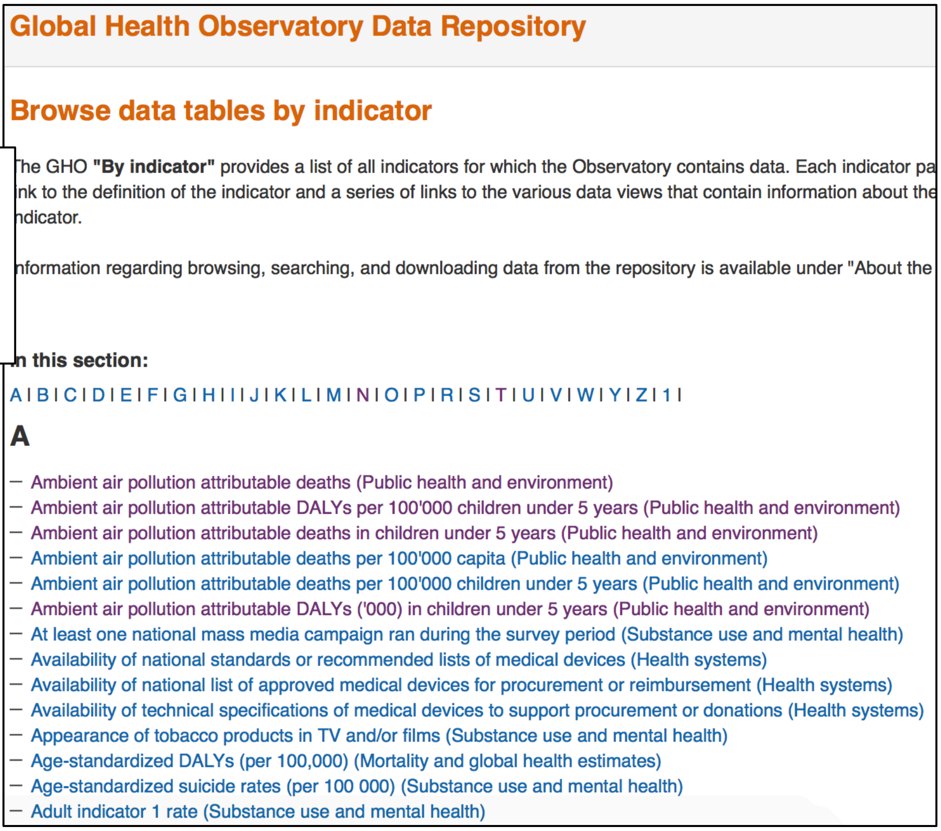
In Step 1, we gather together indicators and results from a variety of organizations. Although these indicators may be measuring the same thing, we will need to perform some intermediary analyses to find out if this is the case. For this reason, in Step 2, we create indicator categories—"themes"—which will cluster together indicators of similar types. In Step 3, we use Machine Learning algorithms to pull our raw list of indicators into the appropriate groups. This makes our overall indicator list (numbering already in the thousands) much more manageable. Then in Step 4 we can look in each theme for variables that overlap—this is the cross-walking exercise. These overlapped indicators will be documented so that we can assess what calculations or corrections need to be made such that they are directly comparable.

General approach: preview

In the following tables and figures we've provided a preview for what each of these steps can, does, or should look like.

Figure 2: Indicator list example (WHO GHO)

Step 1: Data gathering



Global Health Observatory Data Repository

Browse data tables by indicator

he GHO **"By indicator"** provides a list of all indicators for which the Observatory contains data. Each indicator pa nk to the definition of the indicator and a series of links to the various data views that contain information about the ndicator.

nformation regarding browsing, searching, and downloading data from the repository is available under "About the

n this section:

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | R | S | T | U | V | W | Y | Z | 1 |

A

- Ambient air pollution attributable deaths (Public health and environment)
- Ambient air pollution attributable DALYs per 100'000 children under 5 years (Public health and environment)
- Ambient air pollution attributable deaths in children under 5 years (Public health and environment)
- Ambient air pollution attributable deaths per 100'000 capita (Public health and environment)
- Ambient air pollution attributable deaths per 100'000 children under 5 years (Public health and environment)
- Ambient air pollution attributable DALYs ('000) in children under 5 years (Public health and environment)
- At least one national mass media campaign ran during the survey period (Substance use and mental health)
- Availability of national standards or recommended lists of medical devices (Health systems)
- Availability of national list of approved medical devices for procurement or reimbursement (Health systems)
- Availability of technical specifications of medical devices to support procurement or donations (Health systems)
- Appearance of tobacco products in TV and/or films (Substance use and mental health)
- Age-standardized DALYs (per 100,000) (Mortality and global health estimates)
- Age-standardized suicide rates (per 100 000) (Substance use and mental health)
- Adult indicator 1 rate (Substance use and mental health)

This is an example list of indicators from the WHO from their Global Health Observatory (GHO). It is indicators like these that we have organized into datasets. Next we will look at some themes from the WHO's Top 100 health indicators.

Step 2: Creating themes

Table 1: Top 100 health indicators (sample)

Indicator name	**Theme**	**Subtheme**
Life expectancy at birth	Health status	Mortality by age and sex
Adult mortality rate between 15 and 60 years of age	Health status	Mortality by age and sex

Under-five mortality rate	Health status	Mortality by age and sex
Antenatal care coverage	Service coverage	Reproductive, maternal, newborn, child and adolescent
Births attended by skilled health personnel	Service coverage	Reproductive, maternal, newborn, child and adolescent
HIV care coverage	Service coverage	HIV
Antiretroviral therapy (ART) coverage	Service coverage	HIV
HIV viral load suppression	Service coverage	HIV

This table (Table 1) shows a sample of the Top 100 health indicators, which is a list compiled by the WHO and other development organizations of the most important health data to gather in the field. We plan to use themes like these to classify our data.

Table 2: WHO/ UNICEF classification using the Top 100

Indicator	**Actual theme**	**Algorithm #1**	**Algorithm #2**	**Algorithm #3**	**Algorithm #4**
Proportion of all deaths attributable to HIV/AIDS	Mortality by cause	Health information	Mortality by cause	Mortality by cause	Mortality by cause
Percentage of health facilities with the capacity to deliver appropriate care to people living with HIV and AIDS	HIV	HIV	Access	Mortality by cause	HIV
HIV prevalence among pregnant women	Morbidity	Morbidity	Morbidity	Morbidity	Morbidity
Migrants: HIV prevalence	Morbidity	Morbidity	Mental health	Quality and safety of care	Morbidity
Prisoners: HIV Prevalence	Morbidity	Morbidity	Mental health	Quality and safety of care	Morbidity
AIDS-related mortality	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause
Total correct	**—**	**5 (83%)**	**3 (50%)**	**3 (50%)**	**6 (100%)**

Step 3: Classification

This table (Table 2) shows one of our preliminary attempts at classifying raw data. You can see the indicator name in the first column and it's actual theme (according to the Top 100) in the second column. The remaining columns show a variety of Machine Learning algorithms and their attempts to classify the data. Cells in green indicate when a given algorithm made a correct prediction. Already there are certain algorithms that work well given the themes that have been created and the data we have been using (e.g. Algorithm #1 and Algorithm #4).

Step 4: Cross-walking

Table 3: Similar HIV indicators across organizations

Indicator #1	Org. #1	Indicator #2	Org. #2	Distance
TB/HIV mortality rate	Global Fund	HIV mortality rate	WHO	3
HIV prevalence among men	DHS	HIV prevalence	WHO	10
HIV prevalence among women	DHS	HIV prevalence	WHO	12
TB/HIV mortality rate	Global Fund	HIV prevalence	WHO	14
Pregnant women tested for HIV during ANC visit	DHS	Pregnant women tested for HIV, reported number	WHO	16

This final table (Table 3) shows a small list of HIV indicators (i.e. indicators that were classified as "HIV" in the previous step) that are similar across organizations. From here, we plan to research each indicator individually to see how it is calculated—in this way the "HIV prevalence" measured by DHS can be directly compared with "HIV prevalence" as measured by WHO and these statistics can even be aggregated.

Steps in detail

We will now go into each step more in detail and outline the word we've been doing and our plans to move forward.

Step 1: Gathering data

To complete the crosswalk activity, the first step is to gather and categorize types of data. There are two main types of data in which we are interested—indicator data and results data. Often there is considerable overlap between these data types, the main difference usually being the level of aggregation. Results data is at the project or initiative level and is usually found in project reports or final evaluations. Indicator data is at an aggregate level and is usually found in online databases such as the World Development Indicators.

At Step 1, we are not as concerned about data types, although it is important to gather as much metadata information about indicators as possible. This will help us distinguish levels of aggregation later on in the data compilation and comparison process. So far we have been able to compile a fairly large list of indicators (over 4,000) from the following organizations.

1. World Bank HealthStats
2. World Bank Microdata Library
3. World Bank World Development Indicators
4. WHO Global Health Observatory (GHO)
5. WHO Mortality Database
6. Global Fund
7. Global Fund Monitoring and Evaluation Requirements
8. Demographic Health Survey (DHS)
9. USAID Document Experience Clearinghouse
10. Global Health Data Exchange (GHDx)

Step 1: What we need

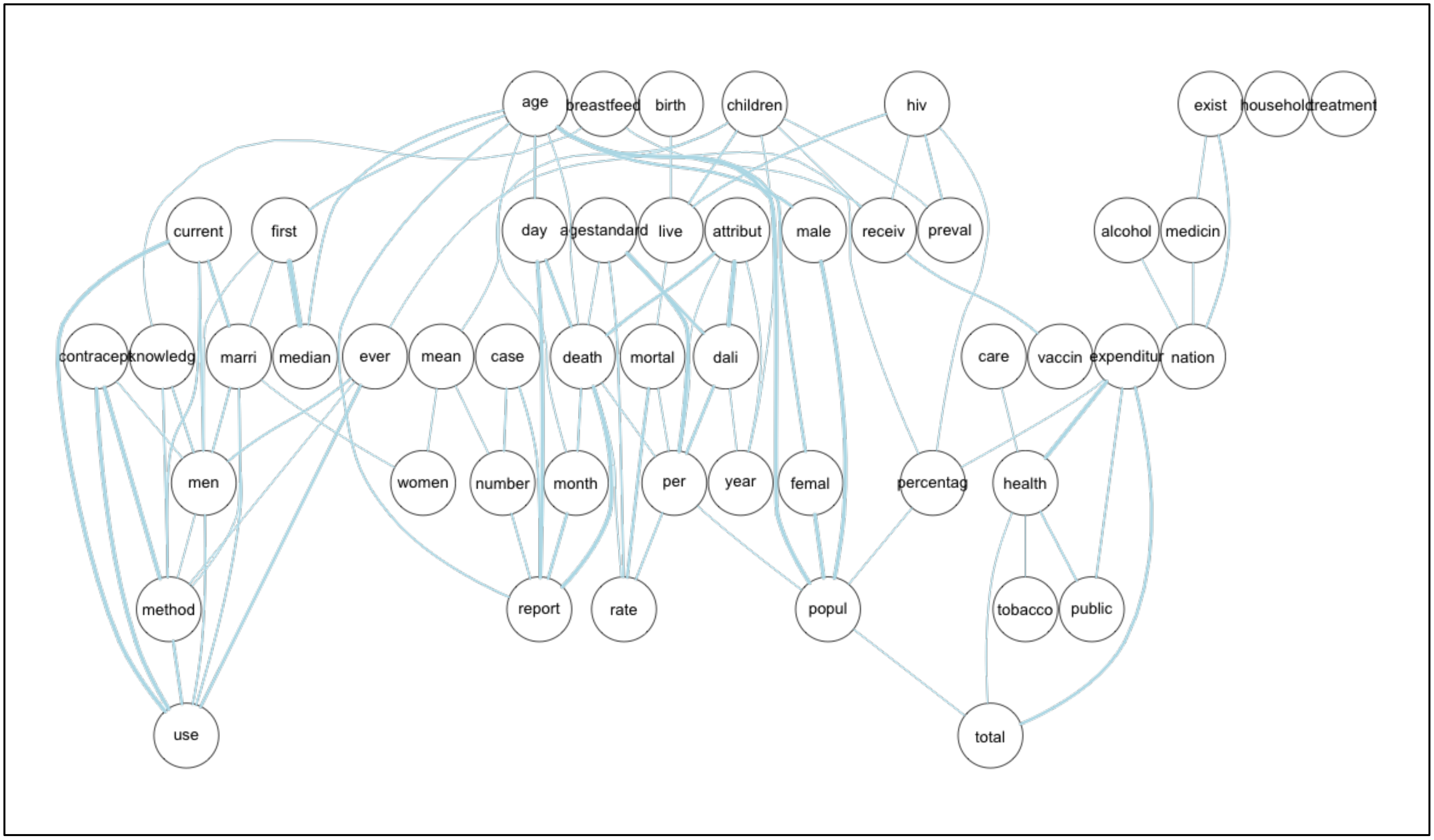
We need all the available data we can get our hands on. More specifically, we need lists of indicators and, ideally, accompanying metadata. Most of this is available online for larger organizations, but we'll need advice on what smaller organizations we should look at and where we can get their data. We also need a reliable way to extract results data, which is mostly contained in PDF reports. For right now, we'll need someone to pull this information out manually.

Step 2: Creating themes and classification schema

There are two main ways to create groups for data organization. The first is to create them "by hand" using knowledge of the field or by finding a standardized list. In the health sector, for example, the WHO, in conjunction with several other aid donors, has created a list of the Top 100 health indicators of interest. This list is already organized into themes and subthemes, which can be used for data organization.

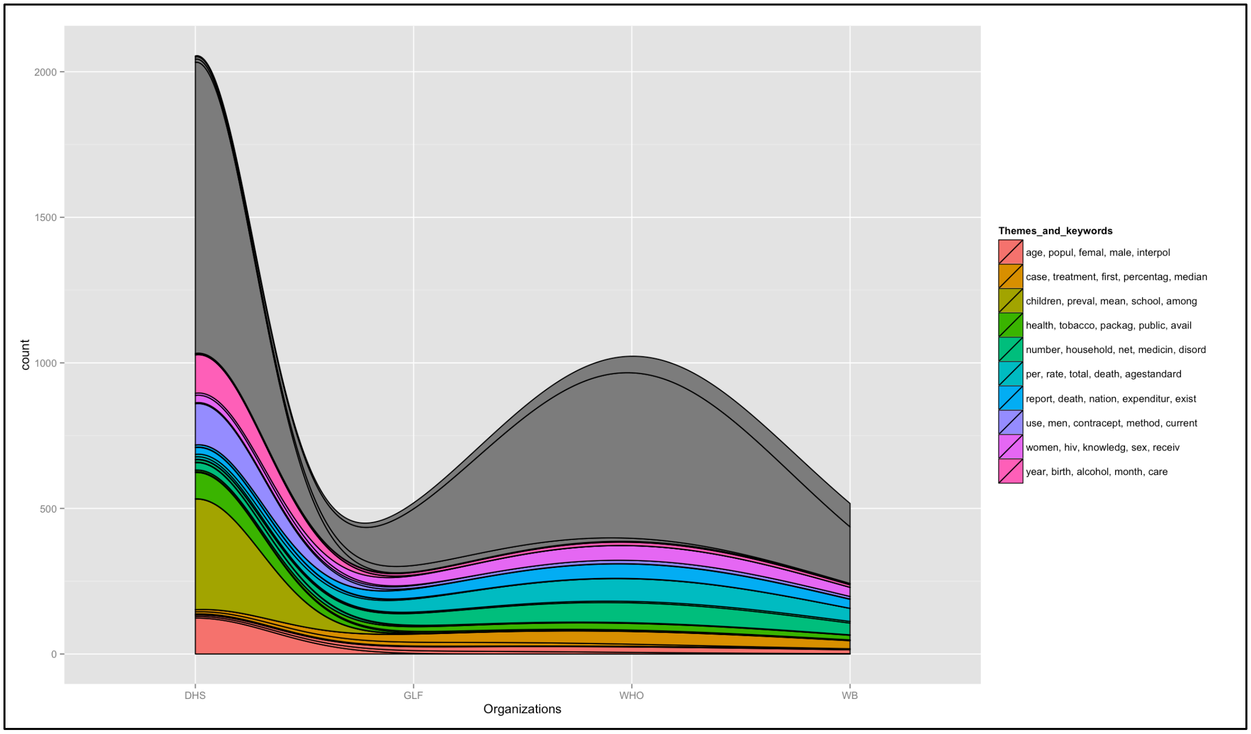
Although a helpful baseline, many of these indicators are highly specific and may miss related variables that could be informative. Thus, a second option is to run classification algorithms to find patterns in the data. This is advantageous since it can be more flexible and use more data, however this can also be a drawback—more data overemphasizes "common" indicators and often neglects more specialized ones such as disease variants (e.g. cervical vs. prostate cancer) or poor health covariates (e.g. insecticide use). Even using heavier weights or normalizing methods (such as the commonly used TF/IDF method) doesn't give us the specificity we would like. The figure below demonstrates this—it shows the most common terms and associations in a group of almost 4,000 indicators (comprised of WHO, DHS, World Bank, and Global Fund data). This figure shows the top 50 most common words, which were words that appeared at least 70 times in the original list. Even when adjusting this aggregation and these associations using TF/IDF weights, uninformative terms such as "current," "nation," and "attribute" still appear.

Figure 3: Association of terms (WHO, DHS, WB, and GF)



Latent Dirichlet Allocation (LDA)[1], which is basically a different way of searching for patterns in word frequencies, produces similar results. The figure below shows the top 10 predicted themes using term frequency weights and LDA analysis. Each theme has a group of 5 keywords.

Figure 4: LDA predicted themes and coverage among donors



You'll notice that LDA is pretty good at producing groups of similar size across organizations, but this isn't actually what we would prefer. Again, although HIV or maternity-related variables may comprise large groups, we'll want to look beyond basic patterns since our end-goal is crosswalking indicators instead of finding overall patterns. In addition, you can see that the majority of both DHS and WHO indicators don't fit in any of these themes.

More metadata can help—i.e. indicator descriptions—but still tends to overemphasize frequencies. With these considerations, it will be most beneficial to use a combined approach. We can use the Top 100 list as a baseline and verify its accuracy and usefulness using word frequency analysis. Additionally, we can update word frequency results using other indicator lists and expertise from people involved in the project. By combining the strengths of each categorization type, we can ensure the formation of detailed groups. This is important for the next phase of our analysis—classifying new indicators based on our categorical list.

Step 2: What we need

We need a solid list of key terms and themes. To get this we'll need advice on lists we're currently using and what they might be missing or overemphasizing. This will especially be the case as we move into agriculture since there is much less consensus on high-priority indicators and key themes.

[1] LDA views each group of words as a "document" and uses a Dirichlet prior distribution to assess content between documents. It uses word frequencies between documents to cluster topics that appear with high levels of correlation.

Step 3: Data classification

Once we have a working list of key themes and indicators, we can start classifying new indicator data. For this we will be using various Machine Learning methods such as Support Vector Machines (SVM) and Neural Networks. We won't go incredibly in-depth as to the statistical specifications of these different algorithms in this report; suffice to say they are methods that "learn" from one set of classifications and then apply these learned rules to new data. This is why a solid list of key indicators is important—the better and more detailed our initial list, the better our algorithms will be at predicting the categories of new indicator lists.

For example, consider the following rough-draft analysis where we use the list of Top 100 health indicators to train our model (i.e. tell it to "learn"). Once we have trained the data, we then classify a new list of indicators—in this case a list of almost 2,000 WHO variables—using seven different algorithms (and TF/ IDF weights). The following table shows a sample of WHO indicators and how they were classified using the different Top 100 lists and algorithms. Cells shaded green are those that were classified correctly.

Table 4: WHO classified using Top 100 (indicator names only)

Indicator	Actual group	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7
Life expectancy	Mortality by age and sex	Morbidity	Mortality by age and sex	Mortality by age and sex	Nutrition	Noncommunicable diseases	Noncommunicable diseases	Mortality by cause
Infant mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Quality and safety of care
Under-five mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Quality and safety of care
Adult mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Environmental risk factors	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Quality and safety of care
Neonatal mortality rate	Mortality by age and sex	Mortality by cause	Mortality by age and sex	Mortality by age and sex	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
TB mortality rate	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
HIV mortality rate	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
Malaria mortality rate	Mortality by cause	Mortality by age and sex	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by age and sex	Quality and safety of care
Suicide rates	Mortality by cause	Mortality by cause	Mortality by cause	Mortality by cause	Morbidity	Morbidity	Morbidity	Mortality by cause
Adolescent fertility rate	Fertility	Noncommunicable diseases	Fertility	Fertility	Noncommunicable diseases	Noncommunicable diseases	Morbidity	Mortality by cause
Total fertility rate	Fertility	Noncommunicable diseases	Fertility	Fertility	Fertility	Fertility	Morbidity	HIV
HIV prevalence	Morbidity	Morbidity	Morbidity	HIV	HIV	Morbidity	Morbidity	Reproductive, maternal, newborn, child and adolescent
Malaria incidence	Morbidity	Morbidity	Morbidity	Environmental risk factors	Morbidity	Morbidity	Noncommunicable diseases	Reproductive, maternal, newborn, child and adolescent
Ambient air pollution in urban areas	Environmental risk factors	Reproductive, maternal, newborn, child and adolescent	Environmental risk factors	Environmental risk factors	Environmental risk factors	Environmental risk factors	Noncommunicable diseases	Noncommunicable diseases
Total correct	—	**5 (36%)**	**14 (100%)**	**11 (79%)**	**9 (64%)**	**7 (50%)**	**5 (36%)**	**1 (7%)**

Note: Methods are (in order of columns) Support Vector Machines, Maximum Entropy, Boosting, Bagging, Random Forests, Regression Tree, and Neural Nets

Side-note: Fertility rates classified as "noncommunicable diseases" and/ or "morbidity"? Do the machines force us to ask tough ethical questions?

After several different model types, the discussion continues as to the best weights and amount of information—i.e. how much information do we use for training and how much emphasis do we give to overall word frequencies. The challenge is the need to restrict how we train the data (i.e. the original list of themes) such that groups are easily identifiable. However, we want to avoid creating groups that are so specific that they exclude relevant matches. And in terms of metadata vs. no metadata—more information can definitely help with classification, but we do not want to oversaturate the model such that relevant words and topics disappear. It represents a delicate balance, especially because the training (original list) and testing (new lists of unclassified indicators) datasets come from different sources—usually model validity can be tested by parsing the training data into training and testing subsets.

Thus, the plan moving forward is more accuracy in two ways. First, we plan to completely classify new sample datasets by hand so that we can better assess overall model accuracy. And second, once we completely classify a new dataset, we plan to incorporate this into the original training data. This way the training data will continue to pull from more information and new datasets will find partner indicators more quickly and accurately.

Step 3: What we need

For now we just need to keep plugging away at the data work and improving the models. Help with the first two steps—particularly with compiling a good list of themes and keywords—will be invaluable to this process.

Step 4: Data crosswalking

The final step is to organize the data (which has now ideally been classified correctly) so that we can identify individual indicators that are the same or similar across organizations. To do this, we'll likely use edit distance algorithms which will compare indicator names themselves. A quick example of this can be seen in the table below—these variables have already been classified as HIV indicators.

Table 5: Similar indicators based on name

Original indicator name	Donor	Matched indicator	Donor	Edit distance
TB/HIV mortality rate	Global Fund	HIV mortality rate	WHO	3
HIV prevalence among men	DHS	HIV prevalence	WHO	10
HIV prevalence among women	DHS	HIV prevalence	WHO	12
TB/HIV mortality rate	Global Fund	HIV prevalence	WHO	14
Pregnant women tested for HIV during ANC visit	DHS	Pregnant women tested for HIV, reported number	WHO	16

This is where human intervention will likely be the most crucial—once we have this list of similar indicators, researchers will need to document the calculations and corrections that need to be made such that the data itself is comparable. For example, "HIV prevalence" as calculated by the Global Fund may be compiled using aggregated clinical data, and thus may be weighted at district or provincial levels while the same indicator from the DHS may be from household surveys and would thus be weighted individually. So, to compare these statistics we would need to correct for their different weighting schemes. A list of these corrections/ needed calculations would be the ideal output at this stage—this is how the actual cross-walking would take place.

Step 4: What we need

Eventually we'll need research assistants to work within these indicator groups and document all the needed corrections. Once we having a working example, they should be able to pull from that and we can start to make this into a workflow. We'll also have a better idea of how intensive the process is and thus how many people/ much time it would take to create the initial database.

Recap and moving forward

We've covered a lot of ground in this report. In an effort to be more explicit about the way forward and how everyone can be involved in the process, we've made the "what we need" sections into a bullet list below. In addition, we've added names to each item to identify responsibilities and levels of involvement.

What we need/ need to do moving forward:

- More data
    - [Daniel/ intern] Get more health data (with metadata if possible)
        - Priority organizations: National governments, USAID, DFID, and AfDB
    - [Daniel/ intern] Get more agricultural data
        - Priority organizations: World Bank, WFP, FAO, and IFAD
    - [Daniel/ intern] Get a sample of PDFs to work with
        - Priority source: USAID Development Experience Clearinghouse
    - [Daniel/ intern] Extract results data from the PDFs
    - [Daniel/ intern] Organize overall indicator/ results database
- Better creation of themes
    - [Danny] Fine-tune Top 100 (better categories, edited metadata, etc.)
        - [Daniel/ intern] Cross-check with OECD-DAC categories
    - [Susan/ Brian/ others] Run the list/ keywords by Susan and Brian to fine-tune more
    - [Danny] Create a rough draft of agricultural themes
        - [Daniel/ Danny] Run term frequency models on agricultural variables
        - [Daniel/ Danny] Also cross-check with OECD-DAC
        - [Brian/ Susan/ others] Run by Brian and Susan to fine-tune
- Better classifications
    - [Danny] Classify entire list by hand (health)
    - [Danny] Tweak theme list, keywords, and algorithms until list is classified at 85-90%
- Draft of crosswalk
    - [Danny] Work with above list once classified and develop edit distance algorithms
    - [Danny] Find metadata for most-similar indicators
    - [Danny] Create crosswalk draft for most-similar indicators
    - [Danny] Create shortlist of example cross-walked indicators
    - [Danny] Create "what we can do with this" exercise