Social Progress Index^{x}: Data Collection and Maintenance Guidelines

By Petra Krylova (pkrylova@socialprogress.org)

Data collection represents a substantive part of developing an Index. Throughout the process data may be collected at various stages and by several people, it is therefore essential to maintain as standardised processes and records as possible. This note aims to provide guidance and suggest best practices, but it is not exhaustive, and each project might require a different approach. It is however fundamental that all is well documented.

Following these simple guidelines will help everyone navigate through the complex maze of 40+ indicators that usually constitute a Social Progress Index.

1) Structuring data

The Social Progress Index framework offers a useful structuring method for folders, it is very intuitive and easy to navigate. Ideally, data should be stored according to Dimensions and Components, the following acronyms and abbreviations have been used:

- BHN
  - NBMC
  - WS
  - S
  - PS
- FoW
  - ABK
  - AIC
  - HW
  - EQ
- Opp
  - PR
  - PFC
  - Incl
  - AAE

In some cases, where one dataset covers more than one component, either a “general” folder can be added, or file does not need to be saved in a folder. This applies especially for sources such as census, Demographic Health Survey etc.

2) Storing and manipulating data

a) Storing summary information on indicators in one place (see Template 1)

From the very beginning it is useful to keep a summary spreadsheet with key information about considered indicators, irrespective of whether they make it to the “final”. The spreadsheet should be structured according to SPI’s framework, and should include the following information: indicator name, indicator definition, indicator source, time period, indicator source link, inclusion in final index (y/n), reasons for rejection, transformation(modification).

b) Maintaining original data

Calculating a Social Progress Index involves numerous steps in data manipulation. It is important to document these well and **maintain the original file untampered with**, so that it is clear what the original data was, and it is possible to keep coming back to the original data set over the course of Index development. If the original file is overwritten, it becomes challenging later in the process, and sometimes it is not possible anymore to retrieve the original data. This also applies for data that was manually copied from a web source, for example. The original storing file should include a header, or a text window with information on how it was collected, the exact source, and date. The source needs to be explicit, a general website of the World Bank for example, will not help locate the data. File containing original data should be labeled "original". It is also important to **use the original source of data** to the extent possible, rather than using others that have re-published it. For example the World Bank Development Indicators contain a lot of data on Education, but many of these come from UNESCO – always check the original source of data and use that source to download the data and refer to it in all methodological documents.

c) Missing values (please see sheet missing values and imputations)

More often than not a number of data points might be missing. With global Social Progress Index we use regression to derive missing values. With subnational indices, often other approaches are more suitable. These might include using historical values, averaging values for most "alike" units, using a value for a higher-level geography etc. Such manipulations are more easily done in Excel, where individual missing data cells can be overwritten with the appropriate value, however, the method for imputing missing data must be clear and noted in the document. For example, if missing value is imputed by averaging most alike units, the calculation of the average should be noted in the document. Create a new document for imputing missing values.

d) Indicator manipulation and transformation

Very often raw data are somehow transformed, or otherwise manipulated before they reach the final calculation stage. Sometimes it is easier to do such manipulations and transformations in Excel, in other cases, STATA might be more suitable. Any manipulation must be well documented. If manipulation is done in Excel, create a new file, based on the original data set. Maintain all steps of the calculation, not just the final figures. For example, if calculating share of population with tertiary education by using two indicators - total population and total population with tertiary education - maintain both indicators, as well as all formulae and calculations used to arrive to the final value of population with tertiary education. If using Excel, always ensure the calculation formulae are saved in the document so that it's easy to trace back what has been done. For any calculations in STATA, save the do files.

e) Keeping indicator manipulations separate

Unless multiple indicators have the same original source, it is better to keep indicator manipulations separate. If the original source file includes more than one indicator, manipulations with indicators should be done in separate tabs that are labeled accordingly.

3) Bringing it all together and Indicator labeling

Once data manipulation in excel is completed we can start building a complete dataset that will be imported into STATA (or R) and used for index calculation. This data set should be a separate spreadsheet, and will include all considered indicators. To make calculations in STATA easier, indicator labels should include a prefix according to the component where it belongs, for example: nbmc_indicatorname, ws_indicatorname, sh_indicatorname, ps_, k_, i_, hw_, eq_, pr_, pf_, in_, ae_.

What to look out for:

- **Different taxonomy for units of observation**

Very often sources use different taxonomy, spelling or notation for units of observation. These need to be aligned in order to then be able to consolidate all data points in one data set. This should be done in the same spreadsheet as other data modifications. Rather than overwriting the original name, it is better to create a new column that will include the name used for the Index. It is also useful to maintain a summary spreadsheet of all the taxonomy by different sources.