## Gaps Analysis Methodology

A set of 8 explanations, A to H, are used to explain the source of gaps in the data available for the 134 ESG indicators. A Python script served to extract data on most recent values (MAXMRV) for all indicators and assess country level coverage of each indicators. The country level coverage has three categories, small countries (countries below 100.000 people), developed countries, and the rest of countries. The Python script extracted data on the total number of countries for which the specific indicator did not have data, the number of small countries for which the indicator did not have data, the number of developed countries for which the indicator did not have data and the number of other countries with no data for the specific indicator. Another stream of information provided by the Python script was the percentage of small and developed countries for which the indicators did not have data.

Additional research on the metadata together with interviews with data curators within the World Bank were conducted to assign gaps explanations for each indicator.

**Explanation A** covers the indicators that were explicitly dropped or deprecated. The specific indicators have been removed from their original database or they are in a deprecated database (e.g. Africa Development Indicators) or they are available in the World Development Indicators (WDI) Archives. A total of 15 indicators are coded TRUE (1) for explanation A. They belong to two databases, 57 and 11.

**Explanation B** covers indicators that were not specifically dropped or deprecated but that are not actively updated. The cutoff point were indicators that had the most recent value 2014 or lower which indicates that no data for the past 5 years was added to these indicators. A total of 15 indicators were coded TRUE for explanation B.

Any indicator coded TRUE for explanation A or B does not qualify for further explanations as these indicators are either dropped or not actively updated, thus of little to no interest for the ESG work stream.

**Explanation C** covers indicators that have temporal gaps in data availability. It tracks two criteria which lead to gaps in data availability. Firstly, the indicators are published biennially, triennially or more. Secondly, the indicators are derived from underlying data that is at least 2 years old – an indicator like this would have in 2019 data for 2017. The criterion used to code TRUE for explanation C was a most recent value (MRV) between 2015 and 2017, as any indicator with MRV below 2015 was coded already B. A total of 61 indicators were coded TRUE for explanation C.

**Explanation D** covers indicators for which the production cycle of the original source is more frequent than the updates conducted by the WDI team. For example, an indicator is published monthly or quarterly, but the WDI team is updating it on an annual basis. Two indicators only were coded TRUE for explanation D.

**Explanation E** covers the set of indicators for which more recent data exists but licensing and publication rights granted by the indicator’s owner do not allow for publication of more recent data. This is the case for indicators supplied by the International Energy Agency who limits to a certain year, 2014 or 2015 in 2018, the publication of their data. A total of 8 indicators were coded TRUE for explanation E and all are created by the same organization, the International Energy Agency.

**Explanation F** covers indicators that are produced from country level microdata surveys with varying schedules. A total of 30 indicators were coded TRUE for explanation F.

**Explanation G** covers indicators that have missing data for entire classes of countries. An indicator is coded TRUE for G if it has missing data for at least 80% of small countries or at least 80% of developed economies. A total of 36 indicators were coded true for G.

**Explanation H** covers indicators for which data points are available for specific countries, but excluded from publication due to internal decisions.