2 The International Data Infrastructure

Grounded research is research in which theories arise from the data, rather than being imposed on the data. Grounded research is the all-but-unavoidable normal practice in quantitative macro-comparative research (QMCR) because nearly all QMCR, whether positivist or interpretive in its formal epistemology, is built on a common international data infrastructure. The global availability of electronic databases on the internet drives convergence in researchers' choices of source data and even researchers' choices of what problems to address with those data. While it may make sense for researchers to put extraordinary effort into conceptualizing and operationalizing one or two key variables of interest for their studies (as Schedler and Mudde 2010 suggest the best researchers do), it is unrealistic for researchers to take this kind of craft approach to the operationalization of background and control variables. Consequently, much of the data used in QMCR are now drawn from a limited number of standard sources, though often with researchers adding one key variable for one particular time period. In addition to the fact that there is only one world available for study (with all the attendant consequences for hypothesis testing), increasingly there is just one standard global dataset available for studying that world.

Even this is relatively new. For example, the regular, standardized, annual measurement of national income dates back less than a century. In estimating rough per capita national income levels for regions of the world for the millennia before 1914, Maddison (2001) collates figures assembled by cliometricians that are ultimately based on shaky indicators such as diet, household goods, and wages mentioned in employment contracts that happen to have survived in archival collections. Even for rich countries, most direct data for the period before World War I derive from sporadic data collection efforts made in response to specific policy needs (Korzeniewicz et al. 2004:537). In their detailed history and sociology of national income accounting, Korzeniewicz et al. trace the genesis of contemporary national income accounting to economist Simon Kuznets's efforts in 1934 to estimate the economic impact of the stock market crash of 1929 (Kuznets 1934), though Kuznets himself in this paper references earlier work by King (1930), and King cites his own previous work with Mitchell et al. (1921). As Korzeniewicz et al. relate, national income accounting only became regularized during World War II in support of U.S. and U.K. production planning efforts.

In the wake of World War II, attempts were made to standardize national income accounting methods across countries. The first textbook on national income accounting was published shortly after the war (R. Ruggles 1949); in 1953 the United Nations (UN) adopted its first standardized system of national accounts (SNA; Stone 1953). The SNA includes detailed procedures for calculating many key data inputs used in QMCR: not just national income, but also imports, exports, domestic investment, foreign investment, value added by sector, and many other economic statistics. As a result, scholars comparing economic data across countries can be reasonably confident that the reported data represent similar measurement concepts across panels of countries. A major negative consequence of this standardization, however, is the loss of heterogeneity in measurement concepts.

Nearly all QMCR involving economic concepts is, practically speaking, limited to the study of concepts that are defined in the SNA. Other economic concepts (e.g., resource depletion, national wealth, median wages, poverty, and income inequality) are poorly, inconsistently, and irregularly measured. Moreover, even for concepts that are represented in the SNA, perfectly reasonable alternative conceptualizations are often not available. So, for example, Korzeniewicz et al. (2004:543–544) describe how nonmarket forms of production like household labor and subsistence farming are excluded from national income due to early decisions to restrict the SNA to paid labor only. These measurement choices, made in a context in which it is generally impractical or impossible for individual researchers to collect their own data in accord with their own judgments on questions of measurement, have profound consequences from the standpoint of the sociology of knowledge in QMCR. As Korzeniewicz et al. conclude,

> The evolution of the SNA . . . entails the construction of a large-scale information infrastructure. . . . This has produced not merely data, but a new categorical understanding of the world. Information infrastructures (such as the one entailed by the SNA) are constitutive of the usually unexamined backdrop for the categories that enable an ordering of the world. (Korzeniewicz et al. 2004:547)

In parallel with the SNA, other kinds of broadly cross-national data also came to be standardized in the postwar period under the aegises of various UN specialized agencies, especially demographic data (organized by the UN Population Division [UNDP]) and health data (as the World Health Organization [WHO] took over and further developed the earlier International Classification of Diseases). In addition to these worldwide efforts, the Organisation for Economic Co-operation and Development (OECD) was constituted in 1961 as a policy coordination agency for the rich countries of Europe and North America (and later Asia). The OECD collates and reports comparative figures that are much more detailed than those available through the UN agencies, but for a much smaller number of countries (all of which have highly capable national statistical bureaus of their own). Following the rise of the cult of education in the development literature in the 1980s (see Easterly 2001 for a review and critique), both the World Bank and the OECD began to emphasize the collection of cross-national education data, including, in the case of the OECD, the organization of standardized student achievement tests across countries. As a result, education data—but only particular kinds of education data—are now entrenched in the international data infrastructure.

More or less absent from the emerging international data infrastructure are political and cultural data. By comparison with the available economic, demographic, and health data, political and (especially) cultural data have been largely ignored by the major intergovernmental organizations (IGOs) that collect and collate data. In the case of political data, nongovernmental organizations (NGOs) have partially filled the gap, though inevitably allowing their own ideological biases to inform their measurement decisions. (See Munck and Verkuilen 2002 for a comparison of attempts to conceptualize and operationalize democracy.) The regular, systematic collection of cultural data, however, is almost nonexistent, and this is reflected in a general lack of QMCR on culture. The cultural gap is beginning to be partially filled by two major survey efforts targeted at individuals, the World Values Survey (WVS) and the International Social Survey Program (ISSP). Nonetheless, data on potentially important QMCR topics like language use, religious observance, family structure, and living patterns are virtually nonexistent, to say nothing of data on popular culture topics like dress, food, music, art, television, and leisure activities.

A bedrock principle of the international data infrastructure is that organizations collect data in support of their missions, not for the convenience of researchers. This accounts for the overwhelming predominance of economic data in the international data infrastructure, and the distantly trailing but still robust showing of demographic and health data. Cultural data are much less important to the governmental donors that ultimately support IGOs and many NGOs, and political data have the potential to be downright embarrassing. This dynamic explains much or most of the unevenness in the international data infrastructure. Consider the contrasting cases of national income and income inequality data. National income data, a core interest of several well-funded IGOs, are systematically collected every year for most of the world's countries according to a highly developed set of standardized accounting rules, and then are collated at the International Monetary Fund (IMF) by a standing body of full-time, highly paid professionals. Income inequality data, on the other hand, are intermittently collected as a nonstandardized byproduct of censuses and other population surveys for fewer than a third of the world's countries in any given year and are irregularly collated by independent academics, depending on the availability of staff and funding.

Such patterns in the availability of data inevitably drive many decisions of what to study in QMCR. In the remainder of this chapter, three basic elements of the international data infrastructure are laid out by type of data coverage: broad cross-national data, detailed data on rich countries, and individual level data. The sources of broad cross-national data including most or all of the countries of the world are examined first, beginning with today's most comprehensive and widely used data compilation of cross-national data, the World Bank's World Development Indicators (WDI) database. Also included are other, supplementary sources of global country data, including both IGOs and NGOs. Nearly all of these broad sources include data on rich countries as well as poor countries. Second, the sources of more-detailed cross-national data focusing on rich countries are reviewed, beginning with the OECD. Again, supplemental IGO and NGO sources are also included. Next come the two major global social survey efforts, along with compilations of census and polling data, all of which report data on individuals that can be aggregated into country data. Finally, this chapter concludes with a preliminary peek at two emerging forms of data that might inform QMCR in the future: internet-based metadata and systematically collected comparative qualitative data.

SOURCES OF BROADLY CROSS-NATIONAL DATA

A variety of intergovernmental, nongovernmental, and U.S. government sources produce standardized datasets with global or near-global coverage. Though there is no global coordinating body for these efforts, taken together these sources constitute a relatively unified international data infrastructure. The World Bank, in particular, has emerged in the role of a data aggregator, collating data from a wide variety of sources into its omnibus WDI database. For some kinds of data there are multiple competing sources (e.g., demographics), while for other kinds of data coverage is spotty at best (e.g., culture). Some of the major contributors to the international data infrastructure are highlighted in Table 2.1.

Most of the data sources listed in Table 2.1 are primary sources in the sense that they produce data for consumption by professionals and the public, but nearly all of them rely on data reported by national statistical agencies (or their proxies). The WDI is mainly a secondary compilation, though many series in the WDI are generated by the World Bank from primary inputs (e.g., the Atlas series of national income figures converted into U.S. dollars). The CIA World Factbook, on the other hand, is entirely a secondary compilation, except for the inclusion of CIA estimates where primary data are missing. While its completeness makes the World Factbook popular for quick reference, it is rarely used by professional researchers, who typically use primary sources or the WDI instead.

Table 2.1 is relatively complete for generalist IGO sources (though there are many narrow, specialist sources not listed), but only scratches the surface for NGO sources. Some NGO sources are academic or quasi-academic, like the Penn World Table, Polity IV database, and World Income Inequality Database. These are well documented, well vetted, and supported by publications in peer-reviewed journals. Their main shortcoming is that they are not reliably updated year-on-year, due mainly to funding and personnel constraints. Nonacademic NGO sources are generally better-resourced, but often generate little truly original data, instead relying on the repackaging of primary data into compiled indices and rankings. In other cases, they do produce original data series, but they are often of questionable provenance. Nonetheless, some nonacademic NGOs do produce high-quality original data, either through organizational surveys (Transparency International's Corruption Perceptions Index) or by collating widely dispersed data from published sources (the World Economic Forum's Global Gender Gap Report).

Despite the recent proliferation of NGO sources, IGOs still remain the backbone of the international data infrastructure. Nearly all of the UN specialized agencies collect and publish data related to their missions, and the IMF plays a central role in the collection of economic data. It is the World Bank, though, that makes this mountain of raw data accessible and usable to (relative) nonspecialists through the WDI. The existence of the WDI means that researchers using a dozen or more variables do not have to be expert in each of the relevant underlying fields. As a result, the WDI is the one indispensable data resource for QMCR. Researchers may go to the original sources for their key variables of interest, but they are likely to use the WDI for the background control variables that flesh out their statistical models.

The World Development Indicators

The WDI database is released annually by the World Bank late in the year; it contains statistics up through the previous calendar year (e.g., the 2009 WDI contains data for the years 1960–2008 inclusive) and is available free online, along with a supporting

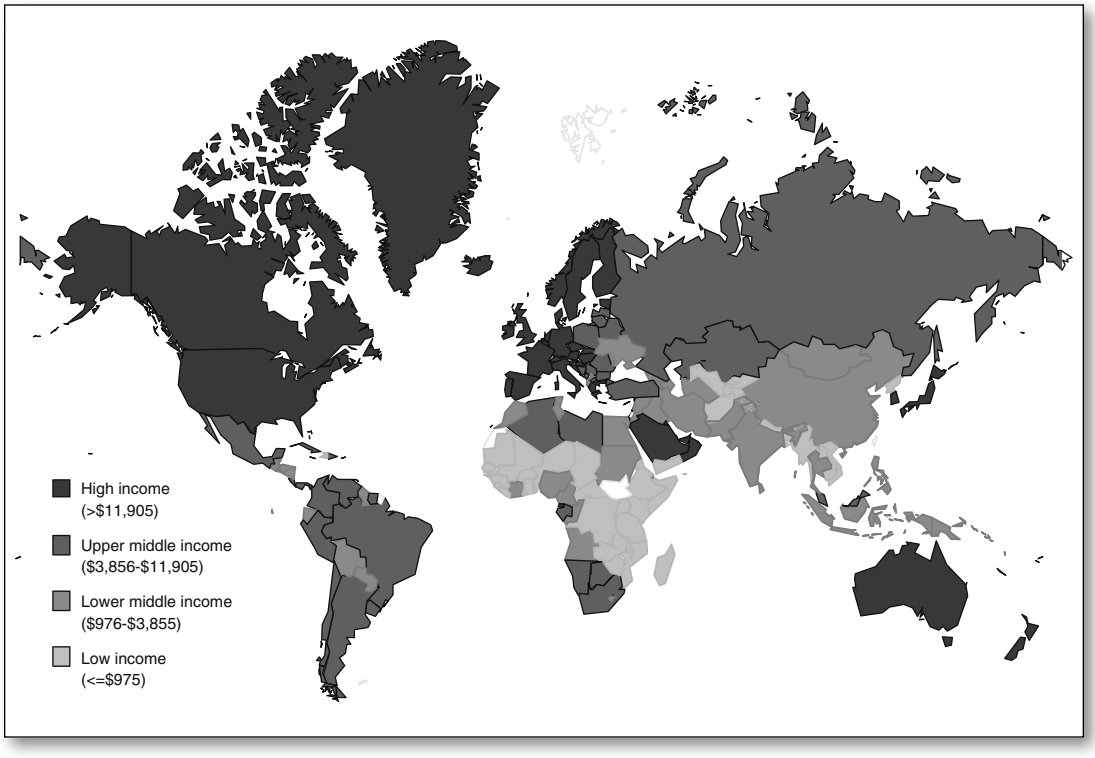
Table 2.1 Major Sources of Broad Cross-National Data (selection)

Intergovernmental Organizations (IGOs)		
P	International Labour Organization	LABORSTA (employment, prices, wages)
P	International Monetary Fund	Balance of Payments Statistics (BOPS)
P	International Monetary Fund	Direction of Trade Statistics (DOTS)
P	International Monetary Fund	Government Finance Statistics (GFS)
P	International Monetary Fund	International Financial Statistics (IFS)
P	United Nations Conference on Trade and Development	World Investment Directory (WID)
P	United Nations Conference on Trade and Development	Trade Analysis and Information System (TRAINS)
P	United Nations Educational, Scientific and Cultural Organization	Institute for Statistics (education data)
P	United Nations Industrial Development Organization	INDSTAT (industrial data including wages)
P	United Nations Population Division	World Population Prospects
P	United Nations Office on Drugs and Crime	Crime Trends Survey (CTS)
S	United Nations Statistical Division	Millennium Development Goals Indicators
P	United Nations Statistical Division	COMTRADE (detailed trade data)
P	World Bank	Global Development Finance (GDF)
P/S	World Bank	World Development Indicators (WDI)
P	World Health Organization	Global InfoBase Online (disease data)
P	World Health Organization	National Health Accounts (NHA)
P	World Health Organization	Statistical Information System (health indicators)
U.S. Governmental Organizations		
P	Census Bureau	International Data Base (IDB)
S	Central Intelligence Agency	The World Factbook
P	Oak Ridge National Laboratory	Carbon Dioxide Information Analysis Center
Nongovernmental Organizations (NGOs)		
P	Center for International Comparisons	Penn World Table (PWT)
P	Center for Systemic Peace	Polity IV Project (governance data)
P	Freedom House	Freedom in the World
P	Reporters Sans Frontiers	Press Freedom Index
P	Transparency International	Corruption Perceptions Index (CPI)
P	United Nations University World Institute for Development Economics Research	World Income Inequality Database (WIID)
P	World Economic Forum	Global Gender Gap Report
P - Primary source (compiled from raw data reported by individual countries)		
S - Secondary sources (compiled mainly from other international sources)		

interpretive book of key tables. The WDI includes data on 869 indicators for up to 210 countries and territories. Ten main areas are covered: education, environment, economic policy and debt, finance, health, infrastructure, labor and social protection, poverty, private sector, and public sector and trade. Nearly every headline summary figure produced by every UN specialized agency is included in the WDI, supplemented of course by many detailed figures and, in many cases, alternative operationalizations of the same indicator (e.g., national income in constant U.S. dollars reported using three distinct methods). The compilation is truly an impressive effort.

The WDI primarily organizes countries by income level (figures in parentheses are numbers of countries and territories so categorized): low (43), lower middle (55), upper middle (46), and high (66). Countries of the world by World Bank income level are mapped in Figure 2.1. All countries except high-income countries are also given one of six regional labels by the World Bank: East Asia and Pacific (23), Europe and Central Asia (24), Latin America and the Caribbean (29), Middle East and North Africa (13), South Asia (8), and Sub-Saharan Africa (47). Since high-income countries are not classified by

Figure 2.1 World Bank Income Groups



Source: Map based on World Bank WDI data.

the region in the WDI, a common practice in QMCR is to create seven regions by lumping all high-income countries into a pseudo region of 66 countries. In addition to these categories, the World Bank labels certain low and lower-middle income countries as highly indebted poor countries (HIPCs). Note that the World Bank regions do not correspond to official UN regions.

The World Bank itself modestly warns readers every year of the limitations of WDI data. In particular, it emphatically cautions against using the WDI for the very purposes that it is used for in QMCR. From the World Bank's official standpoint, the WDI should not be used for statistical modeling in support of policy (though World Bank research staff routinely do just this). The WDI disclaimer grows every year along with the dataset, and is worth reading. The 2012 disclaimer reads, in part:

> Considerable effort has been made to standardize the data, but full comparability cannot be assured, and care must be taken in interpreting the indicators. Many factors affect data availability, comparability, and reliability: statistical systems in many developing economies are still weak; statistical methods, coverage, practices, and definitions differ widely; and cross-country and intertemporal comparisons involve complex technical and conceptual problems that cannot be resolved unequivocally. Data coverage may not be complete because of special circumstances affecting the collection and reporting of data, such as problems stemming from conflicts.
>
> For these reasons, although data are drawn from the sources thought to be most authoritative, they should be construed only as indicating trends and characterizing major differences among economies rather than as offering precise quantitative measures of those differences. (World Bank 2012:xxii)

Such limitations notwithstanding, the WDI is central to nearly all QMCR that is not restricted to rich countries only. Though it may have flaws, in most cases it has fewer obvious flaws (such as misplaced decimal places) than the raw data that underlie it. Anyone reading the extensive disclaimer above is faced with the question: What is the alternative? The comprehensiveness and ease of use of the WDI, augmented by the fact that it incorporates at least some cleaning and auditing of its underlying source data, combine to make the WDI the default source for all QMCR involving broad panels including poor countries. Other sources are used for specific variables of interest, but general background variables are almost always drawn from the WDI when they are available in it.

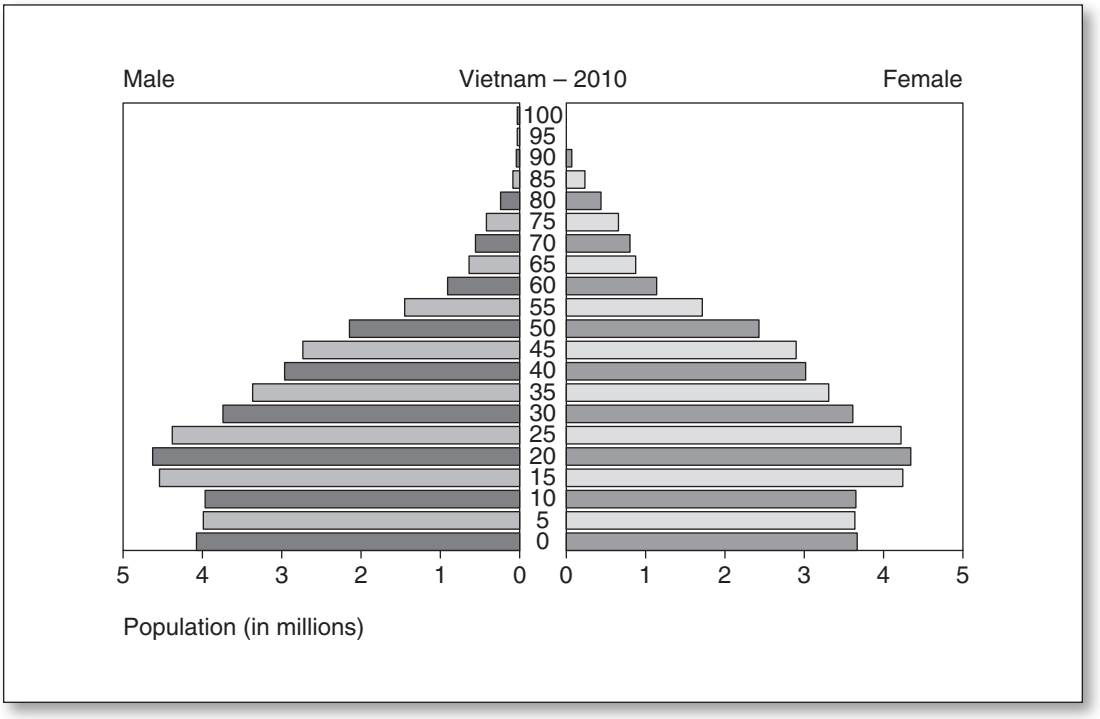
Other Official Sources

In addition to the WDI, the World Bank itself compiles primary source data on development finance (with a particular focus on external debt) and a mix of primary and secondary data on governance and business conditions. The World Bank's annual *Global Development Finance* (GDF) report is representative of the kinds of data produced by the range of UN specialized agencies and the larger UN system (which includes both the World Bank and the IMF). Nearly all these IGOs publish annual reports containing standardized mission-relevant data covering most of the countries of the world. In many cases, these reports also include reference data on population and national income, but most QMCR draws these figures from the WDI instead of relying on the accuracy of data that are incidental to the missions of the agencies involved. The full scope of the available

global data, much of it highly technical, cannot be cataloged here, but most QMCR studies make use of a relatively small number of sources. In fact, many studies use no IGO sources beyond the WDI.

One source that is occasionally used to fill in gaps in WDI data is the CIA World Factbook. As discussed above, it is never appropriate to use the World Factbook in professional research settings as a main data source, but in special cases World Factbook estimates are often used for convenience. This arises most frequently in the case of Taiwan, data for which are suppressed throughout the UN system due to political pressure from China. Use of the World Factbook is a convenient way to fill in missing Taiwan data for studies in which Taiwan is not particularly a research focus, but instead only one of many cases. Similarly, the U.S. Census Bureau International Data Base (IDB) conveniently includes Taiwan estimates alongside those for many sovereign and semisovereign countries and areas that are not included in the World Bank's list of 210 entities. The IDB also has the advantage of reporting projections of population and population attributes 40 years into the future. This includes estimates and forecasts of demographic variables like fertility, life expectancy, and migration. The IDB uses these estimates and forecasts to produce annual population pyramids for the entire period 1950–2050 for a total of 227 entities. A population pyramid for Vietnam in 2010, downloaded from the IDB website, is illustrated in Figure 2.2.

Figure 2.2 Sample Population Pyramid From the U.S. Census Bureau



Source: U.S. Census Bureau IDB.

Though the IDB demographic estimates and projections are slightly more comprehensive, the standard source of cross-national population data is the UNPD. The UNPD figures are the ones that are primarily implemented in the WDI. The differences between UNPD and IDB figures are relatively minor and probably only of interest to researchers whose main focus is demographics. According to Velkoff and Kowal (2006:68), the two organizations "share data sets but use different modeling techniques and assumptions to produce their demographic estimates and projections." It is thus difficult for the nonspecialist to make an informed choice between the two sources. In most practical research settings the selection will make no difference to results, in which case the convenient inclusion of UNPD data in the WDI makes them the default choice except in cases where countries are excluded from the WDI database.

The IMF produces a range of statistics relating to the robustness of countries' financial systems, including the Balance of Payments Statistics (BOPS), Government Finance Statistics (GFS), and International Financial Statistics (IFS) databases. Summary headline figures from these databases are included in the WDI, but for more-detailed data researchers can access these specialized databases. The IMF also publishes the Direction of Trade Statistics (DOTS) database. This is one of the few elements of the international data infrastructure that reports relational (country-to-country) data rather than just compositional (country aggregate) data. Summary trade concentration figures based on DOTS data have been made available for public use by Babones and Farabee-Siers (2012). The UN Commodity Trade Statistics Database (COMTRADE) does the same, only at a far more detailed level, reporting the values of imports and exports of specific commodities between specific pairs of countries. This high level of detail has made the COMTRADE database especially popular for network analyses of global trade (e.g., Mahutga 2006).

A major limitation of all trade data is poor reporting by countries. This is especially true for exports, which aren't as closely monitored as imports. Studies based on compositional country-pair data sometimes use the respective paired-country import figure in place of exports to correct for this, on the logic that one country's imports are another country's exports. Another problem with trade data is that they increasingly no longer represent transactions made at realistic market prices. According to Van den Bossche (2005:9), "approximately two-thirds of all trade takes place within companies," though with no firm data to verify this claim. The proportion is certainly higher than it was in previous decades, and is reputed to still be rising. A figure of 80% seems not unlikely, as multinational corporate consolidation progresses and large retailers increasingly make purchases at the factory gate in low-wage countries and then import the goods so purchased into rich countries as internal transfers. This creates potentially serious distortions in the valuation of international trade, since companies are known to routinely manipulate internal transfer prices so as to book profits in low-tax jurisdictions (Clausing 2002).

Foreign investment data are often used in QMCR in conjunction with trade data. The most detailed foreign investment data come from the World Investment Directory (WID) published by the UN Conference on Trade and Development (UNCTAD). The WID includes relational data on international investment flows, as well as breakdowns of investment by industry using International Standard Industries Classification (ISIC) codes. The WID is the sole primary source of foreign direct investment (FDI) stock data, which forms the basis for foreign capital penetration (PEN) measures. The main shortcoming of the WID is that at finer levels of detail many desired figures are simply unavailable.

In fact, a general principle of working with the international data infrastructure is that the most detailed data sources are also those most likely to be plagued by missing data. Two extraordinarily impressive—but ultimately frustrating—data sources are the LABORSTA database published by the International Labour Organization (ILO) and the INDSTAT database published by the UN Industrial Development Organization's (UNIDO). These UN databases provide highly detailed data on wages, prices, and employment by country by ISIC code. Both are based on the same underlying SNA data reported by member countries to the UN, with the LABORSTA also including other series relating to labor (e.g., unemployment, strikes, work-related injuries), whereas the INDSTAT includes other series that are more relevant to industry (e.g., value added and capital formation). The main problem with both data sources is their unevenness across countries in terms of which industries are covered. This is especially a problem for wages, since average wages cannot properly be compared across countries if the data from different countries are drawn from different industries.

A further complication arises from the fact that both LABORSTA and INDSTAT data are organized according to industry classification rather than occupational classification. This means that the wage figures for any given industry include the earnings of managers and professionals averaged in with those of line workers. This is an artifact of the origins of the data in the SNA, which is fundamentally organized from a business establishment (rather than from a household) point of view. As a result, the wage data in both datasets are ultimately derived from establishment surveys of business and government organizations, rather than from household surveys. The LABORSTA database does include employment, unemployment, and hours worked from household surveys, though not income or earnings data. Most of these underlying household surveys do include questions on income and sources of income, which are collated by the ILO in its Household Income and Expenditure Statistics (HIES) database, but unfortunately these raw data are not standardized across countries, never mind broken down by industry or occupation of respondent.

Other elements of the international data infrastructure are much less developed, and as an incidental result much more straightforward. Education data from the UN Educational, Scientific and Cultural Organization (UNESCO) are relatively comprehensive on a number of detailed indicators, but reach back only to 1999. UNESCO also reports (spotty) data on cultural indicators like cinemas and book publishing. Comprehensive health, disease, and health spending (but not health system) data are published annually by the WHO. The UN Office on Drugs and Crime publishes a range of crime and criminal justice system data based on its recurring (but not annual) Crime Trends Survey (CTS). The International Telecommunication Union (ITU) publishes cross-national data on global internet usage, as well as telephone and mobile phone data; the International Civil Aviation Organization (ICAO) publishes data on air traffic; the Food and Agriculture Organization (FAO) publishes agricultural commodity data; and so on. In addition, the five UN regional commissions (for Africa, Asia and the Pacific, Europe, Latin American and the Caribbean, and Western Asia) publish (mainly economic) data relevant to their regions, though most of these data are derived from the primary sources outlined above.

NGO and Specialist Sources

Like IGOs, many NGOs publish mission-relevant statistics, but relatively few NGOs actually possess serious global data-collection capabilities. As a result, much NGO-produced