

**Data Skills: Capacity Development**  
**TTL: Craig Hammer, DECDG**

**4 Capacity development offerings:**

• <i>Introduction to Open Data</i> curriculum (outline)	Page 1
• <i>Data Literacy</i> curriculum (outline)	Page 2
• <i>Data Analysis</i> curriculum (outline)	Page 6
• <i>Statistics Literacy</i> curriculum (outline)	Page 9

**#1: 'INTRODUCTION TO OPEN DATA' (E-LEARNING & FACE-TO-FACE CAPACITY DEVELOPMENT)**

**Design and Creation of *Introduction to Open Data* Curricula.** The World Bank has developed a series of courses to provide knowledge and skills for both practitioners and users of open data. Each course in the series is designed for a distinct user segment, and provides deep technical skills, extensive examples and case studies, with an emphasis on open data in developing countries.

**Audience.** Each course curriculum is designed to introduce open data to a target audience. Specifically, we offer one course each for data producers, data users, and policymakers. The curriculum is deliberately designed to be portable and extensible, to be useful across country contexts.

**Delivery.** The curriculum comprises 3 separate e-learning courses, which can be delivered as either self-paced courses or as facilitated/group learning. Delivery of each course can be complemented by Training of Trainers workshops to enable further delivery of these courses by local counterparts.

**Course 1: [Open Data for Data Producers](#)** provides a broad overview of Open Data principles and best practices from the standpoint of a data producer, and empowers data managers and technical staff with the background and skills to contribute to the Open Data community. This course is primarily intended for managers and technical staff involved in the production, management, and curation of data, particularly within government ministries. It assumes no prior knowledge of Open Data or specific technical skills. Upon completion, users may take an online assessment test to obtain a completion certificate from the online platform.

**Course 2: [Open Data for Data Users](#)** provides a broad overview of Open Data from a user standpoint, and empowers anyone to take full advantage of Open Data. It is intended for anyone who wants to make better use of Open Data, including ordinary citizens, and assumes no prior knowledge of Open Data or technical skills. Upon completion, users may take an online assessment test to obtain a completion certificate from the online platform.

**Course 3: [Open Data for Policymakers](#)** provides a general overview of Open Data principles and best practices for public policymakers, with focus on the development and implementation of an Open Data program. It is primarily intended for public policymakers in governments that are considering the

establishment or expansion of an Open Data program and assumes no prior knowledge of Open Data or technical skills. This course is much shorter than the other two and provides no certificate option.

## #2: 'DATA LITERACY' CURRICULUM (FOR FACE-TO-FACE CAPACITY DEVELOPMENT)

**Design and Creation of Customized *Data Literacy* Curricula.** This includes curricula for Intro, intermediate, and advanced data literacy capacity development to be customized and delivered, reflective of priorities surfaced from stakeholder consultations. These curricular materials will be handed over (via ToT and further delivery collaboration) to university counterparts to integrate into existing university curricula, as well as to non-government counterparts (civil society, media, NGOs, and think-tanks), to integrate into existing capacity development processes for non-statisticians seeking to build their technical capacity with data analysis and data-driven decision-making.

**Audience.** The 'Data Literacy' curriculum is designed for: Non-government data Users/consumers, including Academia (faculty, undergraduate, post-graduate); selected civil society organizations; selected media; selected private sector and members of the start-up community. The curriculum is deliberately designed to be portable and extensible, to be useful across country contexts.

**Delivery.** The 'Data Literacy' curriculum comprises 14 units, which can be delivered separately (as individual modules, based on demand) or in their entirety (delivered in 50 days over a 1-year period). Delivery of all units is complemented by several Training of Trainers workshops to ensure transfer of technical capacity and local sustainability.

**Unit 1: Introduction to Data and Exploratory Data Analysis.** *Focus is teasing out underlying relationships, trends, and patterns - without resorting to formal statistical notions. The core data concepts that they consider in details include Extremes, Typicality, Variability, and Association.*

- *Module 1:* Introduction to the world of data: definition of data; explanation of data-driven decision making, the lifecycle of data; and introduction to: Types of attributes: Categorical, numerical, ordinal; Visualizing data using dot plots, pie charts, bar charts, and stacked bar charts ; Basic functions, such as dividers, percentage, count, and sum Exploratory data analysis techniques.
- *Module 2:* Finding Insights in Data: case studies of data-driven fact sheets, brainstorming angles for local projects based on sector data.
- *Module 3:* Understanding data formats: an overview of common data formats and ease of use.
- *Module 4 [Optional]: Interactive E-Learning.* The program has developed an interactive learning platform on which learners can interact with (access, analyze, curate, and visualize) data identified / surfaced by the World Bank, using a pedagogy which accommodates both facilitated and self-paced learning. All e-learning curricula, cases and exercises have been developed in English, and will also be precisely translated into Arabic. To access it, visit: <https://sudanebp.tuvalabs.com/> (the e-learning is free: simply sign in using your email address & create a user profile to get started.

**Unit 2: Framing Relevant Questions for Investigation / Making Claims with Data.** In this unit, learners will be introduced to the first step of the data--driven decision making process - i.e. clarifying the issue at hand

by asking relevant questions. They will examine questions to identify the ones that can be answered using data. Further, they will learn to differentiate between questions which seek a deterministic answer and questions that seek an answer based on data that vary. The questions that they ask at this stage will essentially drive the data inquiry cycle.

- *Module 1: Getting Familiar with the Context:* Read the dataset description to get familiar with the context, paying deliberate attention to attributes and their units. Describe the data in a few sentences.
- *Module 2: Assess Validity of Claims:* Study preset visuals to assess the validity of claims made about the data, and explain why some of these claims cannot be made using the data available.
- *Module 3: Studying and Modifying Claims:* Select the claim which can be investigated further using the data. Explore if you can support or refute a claim by viewing group characteristics instead of individual cases
- *Module 4: Modifying & Extending Claims:* Learn to make a statistical claim distinction, that is, if the claim can be investigated by viewing data as an aggregate instead of individual cases.

**Unit 3: Data Sources, Context, and Privacy.** In this unit, learners are introduced to different types of data sources. They learn to pull data from multiple sources, and analyze how context adds meaning to data. They also assess the utility, strengths, and limitations of different data sources.

- *Module 1: Finding Data Online:* Search for data online using advanced search techniques and NGO, think tank, university and government websites.
- *Module 2: Finding Stories in the Facts:* Tell a story with data that has been analyzed and visualized by reliable data sources.
- *Module 3: Strengthening your Data Reliability:* essential questions to ask of each data set to evaluate its integrity.
- *Module 4: Data privacy:* general practices, responsibilities, risks, and key issues.

**Unit 4: Data Collection Methods.** In this unit, learners are introduced to a variety of data collection methods. They consider different sampling designs,

**Unit 5: Data Analysis Using Statistical Methods.** In this unit, learners are introduced to formal statistical measures and concepts. They quantify typicality and variability in data using standard measures of center and variability. Choosing an appropriate statistical measure to summarize data is a key skill that they learn. They also assess the strength of associations between numerical attributes to explore causal relationships.

- *Module 1: Navigating National Databases:* Trying to answer specific questions through national databases including understanding the interface, selecting variables and downloading results
- *Module 2: Interviewing your Data:* basic analysis of a data subset.
- *Module 3: Simplifying Numbers:* Simplify data from large numbers and percentages into ratios, rounded figures and comparisons
- *Module 4: Evaluating Statistics:* evaluate data quality, analyze how statistics are explained, and identify sources and collection methodology of local data.

**Unit 6: Narrating Data Stories.** In this unit, learners collect their findings, and generate data summaries. They interpret the results of their analysis in the context of the data. An important aspect of this module is writing strong, data-backed conclusions and sharing new insights.

- *Module 1:* Reading Charts and Graphs: interpreting data visualized in various chart and graph forms and identifying findings.
- *Module 2:* Creating Compelling Visuals: basic theory of data visualization including matching data and chart types; understanding how data visualization can enhance communication and reports; and how to identify data to include in visualizations.
- *Module 3:* Finding a Visualization for your Story: using a data set from a previous module, identify a story angle, an appropriate title or headline and graphical form and create chart.

**Unit 7: Evidence--Base Decision Making.** In this unit, learners complete the data inquiry cycle, making decisions based on the conclusions they have drawn from their analysis. In this process, they think critically, weighing different perspectives around the issue, and use data as evidence to back their decisions.

- *Module 1:* Crafting Broad Conclusions: Learners organize findings logically, and converging the outcomes into broad conclusions, supporting each conclusion with evidence from the data and their analysis.
- *Module 2:* Making the Shift: Make a shift from hard findings based on data into insights
- *Module 3:* Conclusions in Context: Participants will discuss conclusions as a team, stating its importance in the context of the issue they have at hand
- *Module 4:* Articulating and Sharing Insights: Articulating the insight(s) which emerge out of the conclusions, and sharing them with new audiences, gather reactions, and points of resonance.

### **Unit 8: Data-Driven Storytelling**

- *Module 1:* Story Mapping: Outlining a data story including developing a hypothesis, identifying data sources and structuring your story.
- *Module 2:* Data Expedition: In teams, use pre-selected data sets to tackle a problem, answer a question or work on a project. Each team will have a data-driven product to present at the end of the Unit.
- *Module 3:* Data Expedition
- *Module 4:* Presentations

**Unit 9: Intermediate Data, Data Sources and Exploratory Data Analysis.** *Tease out underlying relationships, trends, and patterns while exploring formal and informal statistical notions.*

- *Module 1:* Review to the world of data: definition of data; explanation of data-driven decision making, the lifecycle of data; and introduction to: Types of attributes: Categorical, numerical,

ordinal; Visualizing data using dot plots, pie charts, bar charts, and stacked bar charts; Basic functions, such as dividers, percentage, count, and sum; Exploratory data analysis techniques.

- *Module 2: Finding Insights in Data:* examine different data sources; hone in on a specific topic; identify public interest angles in indicators, investments and outcomes.
- *Module 3 Finding Data Online:* Search for data online using advanced search techniques and NGO, think tank, university and government websites.
- *Module 4: Finding Stories in the Facts:* Accurately interpret data that has been analyzed and visualized by reliable data sources.

**Unit 10: Making Claims with Data.** In this unit, learners will be introduced to the first step of the data-driven decision making process - i.e. clarifying the issue at hand by asking relevant questions. They will examine questions to identify the ones that can be answered using data. Further, they will learn to differentiate between questions which seek a deterministic answer and questions that seek an answer based on data that vary. The questions that they ask at this stage will essentially drive the data inquiry cycle.

- *Module 1: Assess Validity of Example Claims:* Study preset visualizations to assess the validity of claims made about the data, and explain why some of these claims cannot be made using the data available.
- *Module 2: Studying and Modifying Claims:* Select the claim which can be investigated further using the data. Explore if you can further support or refute a claim using additional data.

**Unit 11: Wrangling with Messy Data.** In this unit, learners study raw data in tabular and other formats, exploring and visualizing the data to identify errors, gaps, and other anomalies.

- *Module 1: Exploring Data through Excel:* Calculate sums, rates, ratios and percentages
- *Module 2: Simplifying Numbers:* Simplify data from large numbers and percentages into ratios, rounded figures and comparisons
- *Module 3: Converting Data to Friendlier Formats:* Scrape data from online sources including converting PDFs to CSVs and Excel files
- *Module 4: Scraping Data from the Web:* Using browser extensions and Google spread sheets to pull in data from web pages.

**Unit 12: Data Analysis Using Statistical Methods.** In this unit, learners are introduced to formal statistical measures and concepts. They quantify typicality and variability in data using standard measures of center and variability. Choosing an appropriate statistical measure to summarize data is a key skill that they learn. They also assess the strength of associations between numerical attributes to explore causal relationships.

- *Module 1: Navigating International Databases:* Navigating the World Bank databank and other international data sites to download data to explore hypotheses, conduct regional comparisons and evaluate public service delivery.
- *Module 2: Inequality Data Analysis using spreadsheets:* Sorting, filtering, percentage change and averages

- *Module 3: Finding News Angles in Charts: Finding insights and communicating the data.*
- *Module 4: Understanding Data: Understand basic statistical concepts including causation and correlation and margin of error.*

**Unit 13: Introduction to Data and Maps.** Introduction to Data Mapping: matching data types with map types, when data should be mapped and when it shouldn't, introduction to different mapping tools.

- *Module 1: Introduction to geographic information systems (GIS) and location data.*
- *Module 2: GIS Data types: Getting a good understanding of the difference between point, line, polygon, raster data types in representing geographic data and when to use which one*
- *Module 3: Data formats: What is the difference between shapefiles, GeoJson, KML and when can which format best be used*
- *Module 4: Tools: In depth hands on session interacting with both freely available online/offline tools using sample data in the different data formats as outlined in module 3. Tools to be used: CartoDB, Fusion Tables, Google Earth, QGIS*

**Unit 14: Intermediate Evidence--Base Decision Making.** In this unit, learners complete the data inquiry cycle, making decisions based on the conclusions they have drawn from their analysis. In this process, they think critically, weighing different perspectives around the issue, and use data as evidence to back their decisions.

- *Module 1: Crafting Broad Conclusions: Learners organize findings logically, and converging the outcomes into broad conclusions, supporting each conclusion with evidence from the data and their analysis.*
- *Module 2: Making the Shift: Make a shift from hard findings based on data into insights*
- *Module 3: Conclusions in Context: Participants will discuss conclusions as a team, stating its importance in the context of the issue they have at hand*
- *Module 4: Articulating and Sharing Insights: Articulating the insight(s) which emerge out of the conclusions, and sharing them with new audiences, gather reactions, and points of resonance.*

### **#3: 'DATA ANALYSIS & DATA MANAGEMENT SKILLS' CURRICULUM (FOR FACE-TO-FACE CAPACITY DEVELOPMENT)**

**Data Analysis & Management Curricula.** This includes curricula for intro, intermediate, and advanced data analysis skills (including data management and more sophisticated data analysis than the above Data Literacy curriculum) and coding skills to be customized and delivered, reflective of priorities surfaced from stakeholder consultations. These curricular materials can also be handed over (via ToT and further delivery collaboration) to university counterparts to integrate into existing university curricula, as well as to non-government counterparts (civil society, media, NGOs, and think-tanks), to integrate into existing capacity development processes for non-statisticians seeking to build their technical capacity with data analysis and data-driven decision-making.



**Audience.** The ‘Data Analysis & Data Management’ curricula are designed for: beginner/intermediate programmers/coders in particular, though certain of the content may be useful as introductory content for a range of interested participants, such as members of academia; civil society; media; and members of the start-up community. The curriculum is deliberately designed to be portable and extensible, to be useful across country contexts.

**Delivery.** The ‘Data Literacy’ curriculum comprises 11 separate courses, which can be delivered separately (as individual sessions, based on demand) or grouped together as a linked curriculum. Delivery of all units is complemented by several Training of Trainers workshops to ensure transfer of technical capacity and local sustainability.

**Course 1: Data Management: Access, Storage, and Dissemination.** This course will introduce participants to data management principles, as well as where to access vast troves of useful development data from The World Bank. Participants will learn about best practice approaches and mechanisms to acquire data, with guidance on archiving and responsible dissemination. This course will also introduce several data management tools, resources and other services offered by the World Bank’s Development Data Group.

**Course 2: Data visualization fundamentals.** This course is a primer on practical data visualization, to help participants condense data and convey messages more effectively and to enable more effective ‘consumption’ and decision-making using data. In this course, participants will learn the fundamentals of data visualization (including through the example of the World Bank’s own data visualization flagship publication, the *SDG Atlas*), including approaches to visual design and reproducible visualizations, and receive training on several data visualization tools, including Datawrapper and R ggplot.

**Course 3: An introduction to R’s Shiny package and its use: Apps for Development.** R is a powerful statistical programming language and one of the central programming languages in (official) statistics and data science. It is open source, and available for different OS platforms (i.e. Mac OS, Linux, Windows). This course will expose participants to R’s Shiny package: an elegant and powerful web framework for building web applications. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge. By using Shiny, it is possible to transform your R code into interactive (web) applications, which other users can run, without any knowledge of R itself, and as such it allows you to deploy cutting edge methodological approaches even to low skill environments. This 90-minute course will give you a short introduction to Shiny, and an overview of its capabilities. This course is suitable for R users and non-R users who are interested in the possibilities provided by this package.

**Course 4: Mobile Phone Data 101.** The objective of this course is to introduce participants to call detail record (“CDR”) data and how to analyze it using Python. By the end of the course, participants will have knowledge about the nature of CDR data, how to undertake basic analysis of CDR data, and to understand the kinds of insights that CDR data can enable.

**Course 5: Introduction to Python.** This course will introduce participants to Python, which is a general-purpose programming language used for data science. In particular, this course will walk participants through the nuts-and-bolts of Python for machine learning, learn the steps to build and deploy a predictive tool, as well as explore machine learning concepts (including the basics of training and test data, identifying important features for prediction, and model evaluation metrics).

**Course 6: Introduction to GitHub.** This course will help participants understand, use, and join GitHub, the world's largest community of developers to discover and share better, reproduceable analytics. Participants will learn to use GitHub to manage and share code for World Bank projects, including sharing and improving code through the power of the crowd: participants will work with collaborators around the world and learn how to reproduce other World Bank analyses.

**Course 7: Introduction to Survey Implementation and CAPI.** This introductory course will provide participants with a solid overview of the methods and techniques for designing, administering and implementing household surveys. This will cover both surveys administered by a project team, and surveys done through consulting firms. It will also provide an overview Survey Solutions, the World Bank developed software for survey management and data collection in CAPI, CAWI and mixed modes. Participants will learn about the possible applications, standard use scenarios, user roles and responsibilities, and features of the software that help improve data quality and transparency of the survey management. A brief demonstration, where participants may bring their phones/tablets/laptops, will supplement the overview. This course will make available information on the steps needed to create and monitor the survey process from start to finish.

**Course 8: Text as Data.** This course will introduce concepts and methods to extract economic insights from text. Through hands on tutorials, participants will learn the basics in text mining and natural language processing. The tutorials will focus on applications of text analytics for social and economic development.

**Course 9: Introduction to World Bank APIs and how to use them.** Did you know that you can download World Bank data directly from a URL with an API? In this course, participants will learn how to use World Bank APIs and how they can be beneficial, especially for bulk downloads. This course will cover the APIs for the World Development Indicators, the Data Catalog, the Microdata library, and the subnational databases.

**Course 10: Introduction to Open Data.** Open data initiatives have been gaining ground across regions, with more than 500 Open Data catalogs now in operation in countries around the world. This course will enable participants to tap into these free, open resources, as well as learn helpful information about open data (useful for professionals, policymakers, and data users). This course will build participants' technical skills on open data, and enable them to benefit from extensive practical examples of how it is used for impact and transformation, and walk through illustrative case studies.

**Course 11: Understanding Purchasing Power Parities.** This course will train participants to better understand and measure living standards (as well as other economic trends) in real, comparable terms. The course focuses on Purchasing Power Parities (PPPs) data – essentially how price levels across



countries differ by measuring the amount of goods and services that a single unit of a country's currency can buy in another country. This course will share a range of useful applications of PPPs, and how they can help advance projects and research. This introductory course will provide participants with a solid overview of the methods and techniques for designing, administering and implementing household surveys with respect to PPPs.

#### **#4: 'STATISTICS LITERACY' CURRICULUM (FOR FACE-TO-FACE CAPACITY DEVELOPMENT)**

**Customizable *Statistics Literacy* Curricula.** This includes curricula for Intro, intermediate, and advanced statistics capacity development to be customized and delivered, reflective of priorities surfaced from stakeholder consultations. These curricular materials can be handed over (via ToT and further delivery collaboration) to university counterparts to integrate into existing university curricula, as well as to Government counterparts to integrate into existing on-boarding processes for Statistics Bureaus, Ministerial and Planning Department statisticians.

**Audience.** The 'Statistics Literacy' curriculum is designed for: (1) government decision-makers working in the generation, analysis and/or, public dissemination of official data; and (2) university faculty teaching statistics. The curriculum is deliberately designed to be portable and extensible, to be useful across country contexts.

**Delivery.** The 'Statistics Literacy' curriculum comprises **8 units**, which can be delivered separately (as individual modules, based on demand) or in their entirety (delivered in 50 days over a 1-year period). Delivery of all units is complemented by several Training of Trainers workshops to ensure transfer of technical capacity and local sustainability.

##### **Unit 1: Launch, Context-setting and Data Analysis with Excel**

- *Assessment:* Participants' statistical knowledge and skills are assessed with a quiz, including: short conceptual problems with multiple-choice responses (that do not require statistical software); short conceptual questions that require a short fill-in the-blank answer (again, no need for statistical software); and a data analysis exercise, that will require the use of a statistical package of their choice (the exercise and data are package-free).
- *Module 1:* Excel course: data management and simple analysis and plots using Excel 2013
- *Module 2:* Introduction to the e-learning platform and course materials repository

##### **Unit 2: The Basics – Descriptive Statistics**

- *Module 1:* Summarizing the "Center", "Wobble", and "Shape" of Data
- *Module 2:* Measuring and Managing Uncertainty
- *Module 3:* The Normal Distribution (among others)

##### **Unit 3: Estimation and Data Management**

- *Module 1:* Large random samples
- *Module 2:* Using a (large) sample mean to estimate an unknown population mean
- *Module 3:* Small samples – what happens when what you've got just isn't enough?

**Unit 4: Introduction to Statistical Sampling:**

- *Module 1:* Conditions necessary to draw a representative sample
- *Module 2:* Other types of sampling processes and their disadvantages.
- *Module 3:* What are sampling weights, and how do I use them?
- *Module 4:* Data quality assurance and cleaning.
- *Module 5:* Basic elements of statistical sampling design: stratification, clustering and multi-stage sampling.
- *Module 6:* Collecting data over time.
- *Module 7:* Panel surveys.
- *Module 8:* Sample size: why is this important?
- *Module 9:* Negotiating a sample size on panel surveys.

**Unit 5: Testing statistical hypothesis:**

- *Module 1:* A Better Approach to Making Decisions and Inferences Using a Single Sample
- *Module 2:* Hypothesis Testing with Two Samples: Comparing One Sample Mean to Another
- *Module 3:* Logic behind statistical hypothesis testing.
- *Module 4:* The role of hypothesis testing in evidence based decision making.
- *Module 5:* Examples of hypothesis tests that can be useful.
- *Module 6:* Using statistical models (multiple regression with continuous or categorical explanatory variables) to assess complex associations.

**Unit 6: Time Series Analysis:**

- *Module 1:* Time Series Data and Basic Forecasting: Trends and Smoothing
- *Module 2:* Decomposing time series
- *Module 3:* Use of indicators over time: strengths, weaknesses and necessary conditions (for impact assessment).

**Unit 7: Understanding associations and relationships:**

- *Module 1:* Understanding association between variables of different types: continuous, and categorical
- *Module 2:* Describing association between variables
- *Module 3:* Assessing the strength of the association between variables: tables, graphs, hypothesis tests to test these types of relationships.

**Unit 8: The National Statistical System and Your Place In It:**

- *Module 1:* Context for evidence based policy-making, the challenges and the benefits.
- *Module 2:* Institutions in charge of the production of national statistics
- *Module 3:* The legal framework that supports the generation of statistics
- *Module 4:* The National Statistical System (NSS) in Sudan
- *Module 5:* Codes of conduct involved in statistics (including the UN Fundamental Principles of Official Statistics, etc).

- *Module 6:* The technical and policy-making challenges of statistics generation and analysis in fragile states
- *Module 7:* International Indicators, national statistics, purpose-build project level indicators.
- *Module 8:* Producing and using official statistics
  - Strengths and weaknesses common to official statistics.
  - Ad hoc data collections
  - M&E surveys
  - Qualitative data
  - Sources of errors in statistics.
  - Guidance to assess the quality of statistical information.
  - The place, purposes and design of other non-governmental data collections – e.g. impact evaluation and ad hoc surveys.
  - Tracking indicators over time.
  - Administrative data.
  - Attribution of effects due to specific interventions
  - Statistical standards and importance of comparability over time and space.

**CONTINUOUS: Participant mentoring and support.** Participants will also be given individual support on their projects (mini dissertations) which will be the final component of the training course.