BIG DATA AND THE SUSTAINABLE DEVELOPMENT GOALS:

Innovations and Partnerships to Support National Monitoring and Reporting

TRENDS
Thematic Research Network on Data and Statistics

Partners for Review

JULY
2021

Table Of Contents

List of Tables

List of Boxes

List of Figures

List of Abbreviations

ABS	Australian Bureau of Statistics
AIS	Automatic Identification System (marine shipping data)
API	Application Programming Interface
BMGF	Bill and Melinda Gates Foundation
C4DC	Contracts for Data Collaboration
CBS	Statistics Netherlands
CDRs	Call Detail Records
CEOS	Committee on Earth Observation Satellites
CEPEI	Centro de Pensamiento Estratégico Internacional
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station
CPI	Consumer Price Index
DANE	Departamento Administrativo Nacional de Estadística (Colombia)
DE	Digital Earth
DMSP-OLS	Defence Meteorological Satellite Program - Operational Linescan System
DNB	Day-Night Band
DSA	Data Sharing Agreements
EO	Earth Observation
ESA	European Space Agency
EU	European Union
FCDO	Foreign, Commonwealth and Development Office
GBD	Global Burden of Disease
GBMF	Gordon and Betty Moore Foundation
GDP	Gross Domestic Product
GEE	Google Earth Engine
GEF	Global Environment Facility
GEO	Group on Earth Observations
GHSL	Global Human Settlement Layer
GIS	Geographic Information System
GPSDD	Global Partnership for Sustainable Development Data
GPW	Gridded Population of the World
GSS	Ghana Statistical Service
HLG-MOS	High-Level Group for the Modernization of Official Statistics
HLPF	High-Level Political Forum
IIASA	International Institute of Applied Systems Analysis
INEGI	National Institute of Statistics and Geography (Mexico)
ISTAT	Italian National Institute of Statistics
JAXA	Japanese Space Agency
EC-JRC	European Commission Joint Research Centre
MoU	Memorandum of Understanding
NASA	National Aeronautics and Space Administration
NGO	Non-Government Organization
NOAA	National Oceanic and Atmospheric Administration
NSO	National Statistical Office
NSS	National Statistical System
NTL	Night-time lights (satellite imagery)
ODC	Open Data Cube
OECD	Organization for Economic Cooperation and Development
OSM	OpenStreetMap
P4R	Partners for Review
PSA	Philippine Statistics Authority
SDGs	Sustainable Development Goals
SDSN	Sustainable Development Solutions Network
STAC	Spatial Temporal Asset Catalog
TReNDS	Thematic Network on Data and Statistics (SDSN)
UK-ONS	United Kingdom Office for National Statistics
UNCCD	United Nations Convention to Combat Desertification
UNECE	United Nations Economic Commission for Europe
UNEP	United Nations Environment Programme
UNSD	United Nations Statistics Division
VIIRS	Visible Infrared Imaging Radiometer Suite
VNR	Voluntary National Review
WRI	World Resource Institute
WSF	World Settlement Footprint

Acknowledgements

This report was prepared by Dr Cameron Allen, Grant Cameron and Hayden Dahmm from SDSN TReNDS with financial support and oversight from Partners for Review. We would like to thank a number of experts who were interviewed during the preparation of the report for their time and insights, namely: Fredy Rodriguez (Centro de Pensamiento Estratégico Internacional), Lisa Bersales (Philippines Statistics Authority), Omar Seidu and Rachel Bowers (Ghana Statistical Service), Monica Scannapieco (Italian Institute of Statistics), Mattia Marconcini (German Aerospace Centre), Tim Harris (UK Foreign, Commonwealth and Development Office Data Science Hub), Graciela Metternicht (UNSW, Global Environment Facility Scientific and Technical Advisory Panel), Giacomo Falchetta (Fondazione Eni Enrico Mattei). Information from these interviews has been synthesized with information from other sources throughout the paper. Any errors or omissions are the responsibility of the authors.

Initiated in 2016 on behalf of the German Federal Ministry for Economic Cooperation and Development and the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety and implemented by GIZ, the Partners for Review (P4R) network's objective is to help foster robust follow-up and review mechanisms at all levels and to ultimately contribute to achieving the Sustainable Development Goals (SDGs).

Executive Summary

The Sustainable Development Goals (SDGs) present an unprecedented monitoring challenge for governments and National Statistical Offices (NSOs). As a result of increasing demands for data from users, declining budgets, and rising data collection costs, interest has grown in harnessing data from new partners in the national "data ecosystem." The potential for "big data" to support SDG monitoring has incited considerable enthusiasm with many emerging experiences and use cases which underscore the need for increased collaboration and partnership. However, there is poor information-sharing on how partnerships can support national SDG monitoring.

This paper provides guidance for countries in regard to leveraging partnerships to harness big data to support national SDGs monitoring. It synthesizes recent experiences from countries who have used partnerships to harness big data as well as the latest research collaborations that are deriving new, innovative datasets from big data to support SDGs monitoring.

National experience in using big data for statistical production has matured over the past decade, with many new applications by countries for the purposes of official statistics and to support decision making. There are also several recent examples of countries experimenting with big data sources to support national SDG monitoring. Within the broader data ecosystem, many new innovations are also occurring from different actors including the research and academic community, international organizations, geoscience and space agencies, the private sector, and civil society organizations.

Given the pressing need to fill data gaps to support SDGs monitoring, it will be important to harness all promising innovations across the data ecosystem. However, this must be done in a way that ensures national ownership and supports NSOs in their central coordinating role. It is therefore important to ensure emerging global datasets derived from big data that can be used for SDGs monitoring, are developed alongside national capabilities and information systems, and that they do not aim to replace conventional methods.

While considerable opportunities exist for partners to support NSOs and governments with their national SDG reporting obligations by providing the necessary data sources, skills, infrastructure, knowledge, and financing to deliver big data projects, a number of challenges and barriers prevent effective partnerships and applications. These include; a lack of awareness and difficulties to initiating projects, legal and privacy issues, and technical capability and methodological challenges. Governments and NSOs are at different stages in addressing these challenges and their maturity in deriving value from big data, which influences the type of partnerships that are most useful. Due to these differences, there is no one-size-fits-all solution and partners can play many different roles.

In addition to data providers, NSOs may need to partner with technical service providers or technology partners to gain access to capabilities if they are not available in-house. Knowledge brokers or convening organizations may be also needed to raise awareness and build trust. Additional funding is often needed to get big data projects off the ground, which can benefit from an executive sponsor to champion a business case for internal funding allocations, or an external donor willing to provide funding. Finally, a legal or data privacy partner to assist in navigating the regulatory landscape is another important partner that is often overlooked.

Some NSOs have yet to commence their big data journey, while others are quite advanced. The maturity some have achieved can be defined across three dimensions: i) Organizational characteristics: big data awareness, literacy, and support for big data use cases; ii) Technical capabilities: both human and infrastructure; and iii) Governance: the legal and regulatory context.

Big data maturity across these dimensions can also be categorized in one of five stages: nascent, pre-adoption, early implementation, proficient, and mature. As NSOs move through these stages, they require different partners and partnership models to gain greater value from their investments.

This maturity model highlights a range of key areas where NSOs and their partners can target their efforts, including developing big data awareness and literacy, human capabilities and infrastructure, and governance and regulatory frameworks. During the journey from nascent maturity through to proficiency, NSOs will work with and benefit from a range of partners and partnership models.

Drawing from recent experience, this report puts forth a series of practical steps that NSOs can take to develop their big data maturity and identifies the partners who can help along the journey.

1. Introduction

The adoption of the 2030 Agenda and the Sustainable Development Goals (SDGs) has driven the demand for more evidence-based reviews and progress reporting. A key enabler for reviewing progress on the 169 SDG targets and 231 indicators is the collection of timely, accurate, and comparable data in a broad range of economic, social, and environmental policy domains. However, as of December 2020, more than 40% of the SDG indicators are not regularly produced (IAEG-SDGs, 2020), and available datasets are often out-of-date, resulting in a lack of data needed for effective monitoring and implementation (See et al., 2018).

National governments are central to monitoring and reporting on the SDGs (United Nations General Assembly, 2017), with National Statistical Offices (NSOs) playing a pivotal role in data collection, coordination, validation, and quality assurance. And **with NSOs facing increasing demands for data from users, declining budgets, and rising data collection costs (Tam and Van Halderen, 2020), interest has grown in harnessing data from new actors in the national 'data ecosystem'** (Cazarez-Grageda and Zougbede, 2019). This broader ecosystem includes partners from NSOs, line ministries, civil society organizations, academia, and the private sector contributing new data sources, knowledge and skills, or computational infrastructure to support national monitoring of the SDGs.

In recent years, the potential for "big data" to support SDG monitoring has incited both enthusiasm and debate relating to its benefits and risks. **Big data offers myriad benefits** from increasing coverage, and efficiencies due to cost reductions and increased timeliness. H**owever, these benefits come with a number of challenges and risks**, including methodological problems, skills and capability gaps, sustainability and access issues, and data privacy and security concerns. Despite these challenges, there is a general consensus that the statistical community can greatly benefit from big data (Kitchin, 2015, Daas et al., 2015, Struijs et al., 2014, Tam and Clarke, 2015, Florescu et al., 2014, Reimsbach-Kounatze, 2015, Scannapieco et al., 2013, Radermacher, 2018). This has been demonstrated during the COVID-19 pandemic which has disrupted traditional data collection efforts, placed additional pressures on NSOs, and resulted in new partnerships to leverage alternative data sources to bridge data gaps (UNSD and World Bank, 2020).

There are many examples of NSOs working with big data to support the production of national statistics (UNSD and UNECE, 2015, UNESCAP, 2020), **as well as recent instances of countries experimenting with big data sources to support national SDG monitoring** (UNESCAP, 2021). These projects often involve a range of actors from government, civil society, the private sector, and academia taking on a number of different roles. They also draw upon new methods and innovations made by collaborative research teams globally in deriving new datasets from a diverse range of big data sources that can support or complement national SDG monitoring efforts. These emerging experiences underscore the need for increased collaboration and partnership to effectively harness big data for the SDGs. **However, there is poor information-sharing on how partnerships can support national SDG monitoring.**

In this context, this paper synthesizes findings from a review of country experiences and emerging innovations in the use of big data for SDG monitoring. The paper draws on interviews with government representatives and experts across the data sector, a systematic literature review of the latest research innovations in the use of big data for SDG monitoring, as well as a broader review of the grey literature on big data partnership processes. It aims to explore recent experience and lessons **learned by countries in harnessing big data, and provide guidance on how partnerships can support national SDGs monitoring**.

The paper is structured as follows. **Section 2** clarifies concepts and definitions and outlines the value proposition of big data for SDG monitoring. **Section 3** reviews big data types and sources, countries' recent experiences using big data for SDG monitoring, and presents the results from the systematic review on emerging innovations in deriving SDG indicators from big data sources. **Section 4** reviews the big data partnership landscape, discusses the types of partnerships and roles of partners, and proposes a big data maturity and partnerships model for NSOs. Finally, **Section 5** outlines practical guidance for NSOs on partnership processes for the use of innovative big data sources and methods for SDG monitoring.

2. The Big Data Value Proposition for SDGs Monitoring

2.1. Big Data Definitions and Types

Big data is a term that describes large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information (Gandomi and Haider, 2015). The occurrence of big data is largely a post-millennium phenomenon and has become widespread only as recently as 2011 (Sivarajah et al., 2017). And its rapid emergence has been enabled by advances in computational power, ubiquitous and mobile computing, networked storage, new forms of database design, new modes of software-mediated communication and transactions, and data analytics that utilize machine learning (Kitchin, 2015).

Big data have a number of distinct attributes that distinguish them from other data sources. These attributes (**Table 2-1**) are important for understanding the contribution of big data to monitoring SDG indicators, including opportunities and risks (MacFeely, 2019).

Table 2-1. THE 10VS OF BIG DATA

(Laney, 2001, Gandomi and Haider, 2015, Hammer et al., 2017, Metternicht et al., 2020, Sivarajah et al., 2017, Li et al., 2020b)

CHARACTERISTIC	DEFINITION
VOLUME	The number of data records, their attributes, and linkages
VELOCITY	The speed at which they are produced, received, processed, and understood
VARIETY	The diversity of data sources, formats, media, and content
VOLATILITY	The changing technology and data storage
VERACITY	The trustworthiness of the origin and availability
VALIDITY	The accuracy, reliability and quality of the data
VALUE	The business value of data collected
VARIABILITY	The meaning of data continues to change
VULNERABILITY	The personal nature of data and the need for privacy and security
VISUALIZATION	The poor scalability and functionality

Official statisticians have been dealing with a diversity of data sources for decades, with many NSOs regularly using administrative data and experimenting with other big data sources, including citizen-generated data and sensor data (Tam and Clarke, 2015). In the context of official statistics, big data sources are often referred to as "non-traditional data sources" (Fraisl et al., 2020). While there is no universal approach to classifying big data types and sources, the United Nations Statistical Commission (2014) describe **the types of big data** as:

- **Administrative data**: Sources arising from the administration of a program, be it governmental or not (e.g., electronic medical records, hospital visits, insurance records, and bank records)
- **Commercial data:** Commercial or transactional sources arising from the transaction between two entities (e.g., credit card and online transactions, scanner data)
- **Sensor data:** Sensor network sources (e.g., satellite imaging, road sensors, and climate sensors)
- **Tracking or mobile data:** Tracking device sources (e.g., tracking data from mobile telephones and the Global Positioning System)
- **Behavioral and opinion data:** Behavioral data sources (e.g., online searches about a product, a service, or online page views) and opinion data sources (e.g., comments on social media)

2.2 Opportunities, Challenges, and Risks of Big Data for Official Statistics

Ultimately, the costs of investing in big data must be carefully weighed against what they might deliver in practical terms (MacFeely, 2019). While statistical data collection has traditionally been based on surveys and questionnaires, it has been increasingly supplemented and replaced by big data sources, such as administrative data sources, which have been integrated into the information architecture of government and have reduced reporting burdens (Struijs et al., 2014). Combining data from different sources has made official statistics more valuable, and efforts have increased to standardize and harmonize these various sources, with NSOs as the principle administrator of the "data value chain" (Kitchin, 2015).

In the context of the SDGs, big data may offer solutions to data deficits where traditional approaches have so far fallen short (MacFeely, 2019). **In national statistical systems (NSS), big data sources have a strong value proposition** as they can serve a range of purposes, with key opportunities including (Florescu et al., 2014, Tam and Clarke, 2015, Klein and Verhulst, 2017):

- Increasing the scope, breadth and quality of statistical insights from existing and new metrics, including the SDGs
- Enabling more timely data products to fill gaps in time series or meet new demands for real-time data products
- Increasing the granularity of existing datasets by enabling small area estimates
- Reducing costs and reporting burdens associated with traditional surveys
- Driving innovation in new methodologies and attracting new talent

Opportunities associated with big data therefore include scope, timeliness, and resolution, as well as the potential for increasing efficiencies in resourcing and compiling statistics and cost reductions. However, the development of big data is also a significant disruptive innovation, which can present several challenges and risks for NSOs (**Table 2-2**). These include methodological problems, legal and regulatory issues, sustainability and access issues, and data privacy, security, and reputational concerns. The opportunities, challenges, and risks of big data will vary considerably depending upon the type of big data (e.g., open-source or proprietary), the legal and regulatory context, and the institutional capacity settings.

Table 2-2. Opportunities, Challenges and Risks of Big Data for Official Statistics (Kitchin, 2015, Struijs et al., 2014, Tam and Clarke, 2015, Radermacher, 2018, Daas et al., 2015, Reimsbach-Kounatze, 2015, Florescu et al., 2014, Scannapieco et al., 2013, Ali et al., 2016).

OPPORTUNITIES	CHALLENGES	RISKS
• Complement, replace, improve, and add to existing datasets • Reduce costs of traditional surveys • Produce more timely outputs • Reduce reporting burden and survey fatigue • Complement and extend micro-level and small area analysis • Improve quality and ground-truthing • Refine existing statistical composition • Easier cross-jurisdictional comparisons • Enhance linking to other datasets • Produce new and better insights with data analytics • Optimize working practices and efficiency gains in production • Redeploy staff to higher value tasks • Greater collaboration with computational science, data science, and data industries • Greater visibility and use of official statistics	• Forming strategic alliances with big data producers • Gaining access to data, associated methodology and metadata • Establishing provenance and lineage of datasets • Legal and regulatory issues and ethical challenges • Establishing suitability for purpose • Establishing dataset quality with respect to veracity (accuracy, fidelity), uncertainty, error, bias, reliability, and calibration • Ensuring security of data • Developing data analytics and modelling capabilities • Institutional change management • Ensuring inter-jurisdictional collaboration and common standards • Processing, storage and transfer of large datasets • Data volatility	• Mission drift • Damage to reputation and losing public trust • Privacy breaches and data security • Inconsistent access and continuity • Resistance of big data providers and populace • Fragmentation of approaches across jurisdictions • Resource constraints and cut-backs • Privatization and competition

3. Big Data Use Cases for SDGs Monitoring

National experience in using big data for statistical production has matured over the past decade. National statistical systems are innovating with big data sources and applications as part of their commitment to modernize their statistical systems, and many use cases for SDGs monitoring are also emerging. **Within the broader data ecosystem, many new innovations are also occurring from different actors** including the research and academic community, international organizations, geoscience and space agencies, the private sector, and civil society organizations. The scale of these research collaborations and partnerships generally extends beyond national borders, using big data sources to derive new regional and global datasets that can also support national SDGs monitoring. **Given the pressing need to fill data gaps to support SDGs monitoring, it will be important to harness all promising innovations across the data ecosystem**. However, this must be done in a way that ensures national ownership and supports NSOs in their central coordinating role. **Bridging the gap between the broader collective of "big data for SDGs" innovators and national statistical systems requires special attention and could deliver considerable opportunities.**

3.1 National Statistical System Innovations in Using Big Data for SDGs Monitoring

The potential for non-traditional and big data sources to support SDGs monitoring is widely acknowledged (Avtar et al., 2020). For example, recent assessments highlight that existing Earth Observation (EO) systems could generate data for 33 SDGs indicators across 14 goals (Kavvada et al., 2020), while citizen-generated data could contribute to one-third of SDGs indicators (Fraisl et al., 2020).

There are also many new applications by countries using big data for the purposes of official statistics and to support decision-making. For example, projects have been under way for a number of years in areas such as web scraping or scanner data for price statistics, social media data used for consumer confidence indexes, mobile phone data for mobility and tourism statistics, and satellite data for agricultural statistics (UNSD and UNECE, 2015, UNESCAP, 2020, Tam and Van Halderen, 2020) (**Box 3-1**). Yet, these advancements have been largely led by well-resourced and advanced national statistical systems in OECD countries, and there is a need to demonstrate and scale up specific applications and broaden their use beyond specialized scientific and expert communities to support decision makers with meaningful indicators.

Box 3-1

Early Leaders in Using Big Data Sources for National Official Statistics

Italian National Institute of Statistics (ISTAT)

ISTAT has been working with big data since 2013, coinciding with regional directives in the European Statistical System Committee to explore big data sources in official statistics (ESSC, 2018, ESSC, 2013). They have a number of big data projects including web intelligence (web scraping of enterprise websites, scraping of prices, scraping job vacancies), use of social media data (including the daily social mood on economy index based on Twitter data), use of public Sentinel 2 satellite imagery and deep learning through an automated pipeline to produce land cover statistics, and experimentation with sensor data, such as mobile network operator data, smart surveys, and Automatic Identification System (AIS) shipping transponder data for maritime statistics.

United Kingdom Office for National Statistics (UK-ONS)

Prompted by the global financial crisis of 2008, the UK-ONS undertook a review of economic statistics and recommended greater use of big data and data science innovation to produce more timely statistics. A Big Data Team in the ONS now has projects exploring web-scraped price data, smart meter data, machine learning for matching addresses, and natural language processing for coding textual survey responses. Published work and code is made publicly available through a code repository. The UK also established a Data Science Campus in 2017 to investigate new data sources, including big data and administrative data, and helped build data science capabilities for the public good. In 2019, a joint team with the Foreign, Commonwealth and Development Office (FCDO) was launched to apply data science for the global public good, which included the development of big data tools, capacity-building in data science for international development professionals, and mentoring of statistical office staff in partner countries (e.g., Rwanda, Ghana, and Kenya). The ONS Data Science Campus also partnered with the United Nations Economic Commission for Europe's (UNECE) High-Level Group for the Modernization of Official Statistics (HLG-MOS) to coordinate the ONS-UNECE Machine Learning Project in 21 pilot countries, with a new 2021 Group recently launched.

Statistics Netherlands (CBS)

In 2015, Statistics Netherlands became the first statistics office in the world to launch official traffic statistics produced with big data (Daas et al., 2016). Building on this early success, they established the Center for Big Data Statistics to enable the faster production of real-time statistics, increase disaggregation of existing statistics, explore new indicators, and to reduce the administrative burden. The Center includes data scientists (specialists with big data), data scouts (who focus on finding new data sources and arranging access), and domain experts (statistical specialists in CBS). The Center's initial scope of activities included the production of social statistics (safety, health, and housing), economic statistics (prices, internet economy, labor market, and energy transition), mobility statistics (traffic flows, crowd control, and tourism), as well as the SDGs and smart cities. Several work programs also lay the foundations upon which reliable big data statistics can be produced. First, data scouting assures that the data needs are identified and made available for the products intended to be produced. Second, ethics and privacy focus on assuring that measures are taken to protect the privacy of the 'units' in the sources and that only the part of the data that is needed is collected, stored, and processed. Third, methodology and data integration focus on applying big data methodologies. Finally, the focus is on ensuring the regular production of big data based on official statistics is possible. A range of partners from universities, government, private sector and international organizations are invited to contribute.

Australian Bureau of Statistics (ABS) and Statistics New Zealand (Stats NZ)	Building on the early success of Statistics Netherlands, NSOs in the Asia-Pacific region have been using non-traditional data for producing official statistics. In particular, Consumer Price Indexes (CPIs) have received considerable attention (UNESCAP, 2020) incorporating scanner and online data into price measurements. Historically, data for CPIs was manually collected directly from retailers via personal visits, phone calls, or manually from websites. These data collection methods resulted in a relatively small sample of products and services and are expensive and burdensome to collect. Non-traditional data sources have helped NSOs deliver more granular CPI series by significantly increasing sample sizes and the geographical coverage of indexes. Scanner data are electronic records of transactions usually where a barcode has been scanned which can be obtained directly from retailers or market research companies. Online data is electronically collected from retailer websites through a process known as web scraping, or data collection from a website's Application Programming Interface (API). ABS started exploring the use of scanner and online data for price statistics around 2011 and introduced scanner data into the official CPI of Australia in 2014. Similarly, the Stats NZ has implemented scanner data for a range of consumer electronics products into the official New Zealand CPI since around 2014, and web-scraped online data from some accommodation and transport services since 2017.

While much of this work has not been driven by national SDGs monitoring needs, the experiences, methods, and lessons learned are still highly relevant. **There are also a number of recent case studies of countries experimenting with big data sources to support national SDG monitoring**. For example, a recent review of countries using non-traditional data sources specifically for SDG monitoring (UNESCAP, 2021) highlights several recent initiatives, with countries using two types of non-traditional data: Earth Observation (EO) and geospatial data (22 countries) and citizen-generated data (1 country). In the wider grey literature, a range of other national experiences are emerging, which are both country-led or supported through global partnerships and initiatives.

For example, **in the Philippines**, an initial mapping of citizen-generated data holdings identified datasets of relevance for monitoring 81 SDG indicators. They have not yet been used to estimate SDG indicators, however (Paris21, 2020). The Philippine Statistics Authority (PSA) has also undertaken research to measure the rural access index (SDG 9.1.1) and relationships with other social and economic indicators, such as poverty rates and regional GDP (Bantang et al., 2020). In addition, the PSA has been working with the Asian Development Bank's Statistics and Data Innovation Unit and World Data Lab to examine the feasibility of poverty mapping (small area estimates) using satellite imagery and associated geospatial data (including night-time lights imagery) (ADB, 2020) (**Box 3-2**). The Philippines' Advanced Science and Technology Institute has also signed a Memorandum of Understanding (MoU) with the PSA to explore the use of artificial intelligence for the next census on agriculture and fisheries, including training in the use of marine shipping (AIS) and satellite data.

Box 3-2

Using Non-traditional Data in the Philippines and Thailand for Small Area Poverty Estimates (ADB, 2020)

In 2017, the Philippine Statistics Authority (PSA), the Asian Development Bank's Statistics and Data Innovation Unit, and World Data Lab began a project aimed at strengthening the capacities of NSOs in the Asia and Pacific region for monitoring the SDGs, including improving the disaggregation of poverty statistics. The project draws inspiration from methods used in academic research studies that use high-resolution satellite imagery, geospatial data, and powerful machine learning algorithms to complement traditional data sources and conventional survey methods. The Philippines and Thailand were chosen because they had existing initiatives to combine household survey and census data to produce more granular estimates, which provided adequate data for training models. Even by using publicly accessible satellite imagery, in which the resolutions were not as fine as those in commercially sourced images, the project produced predictions that aligned with the government-published poverty estimates.

For the purposes of small-scale poverty studies using non-traditional data, the project found NSOs can streamline their resources and avoid substantial upfront costs by capitalizing on publicly accessible satellite imagery, affordable cloud services, and computational tools. However, data sourcing activities on a larger and more meaningful scale might require NSOs to make significant investments in higher-resolution imagery, faster local computation equipment, and the vast volumes of internet bandwidth required to manage high volumes of data. It is also important for NSOs to expand their investments in human capital, such as appropriately skilled data scientists, to collect, process, and analyze geospatial and other non-traditional data, and to integrate new data collection methodologies into their work programs.

In Colombia, the NSO, Departamento Administrativo Nacional de Estadística (DANE), has engaged with the Data For Now initiative to improve the timeliness and granularity of its poverty data. The initiative is led by four core partners - the Global Partnership for Sustainable Development Data (GPSDD), the World Bank, the United Nations Statistics Division (UNSD), and SDSN TReNDS – and supports countries to use innovative sources, technologies, and methods for statistical production. DANE has partnered with Centro de Pensamiento Estratégico Internacional (CEPEI), a technical partner in the region, and the Bogotá Chamber of Commerce to access administrative data to support the monitoring of SDGs 8 and 9. DANE is also using geospatial data and testing methodologies to report on an additional four SDG indicators, including convenient access to public transport (11.2.1), ratio of land consumption rate (11.3.1), open space in cities (11.7.1), and the rural access index (9.1.1). A range of different data sources are being used, including Landsat and Sentinel images, infrastructure and population data (which are processed with the Google Earth Engine (GEE) platform), ArcGIS, and data science software (**Figure 3-1**). DANE is also exploring the use of geospatial datasets (e.g., night-time lights and vegetation index covariates) for small area estimate poverty mapping.

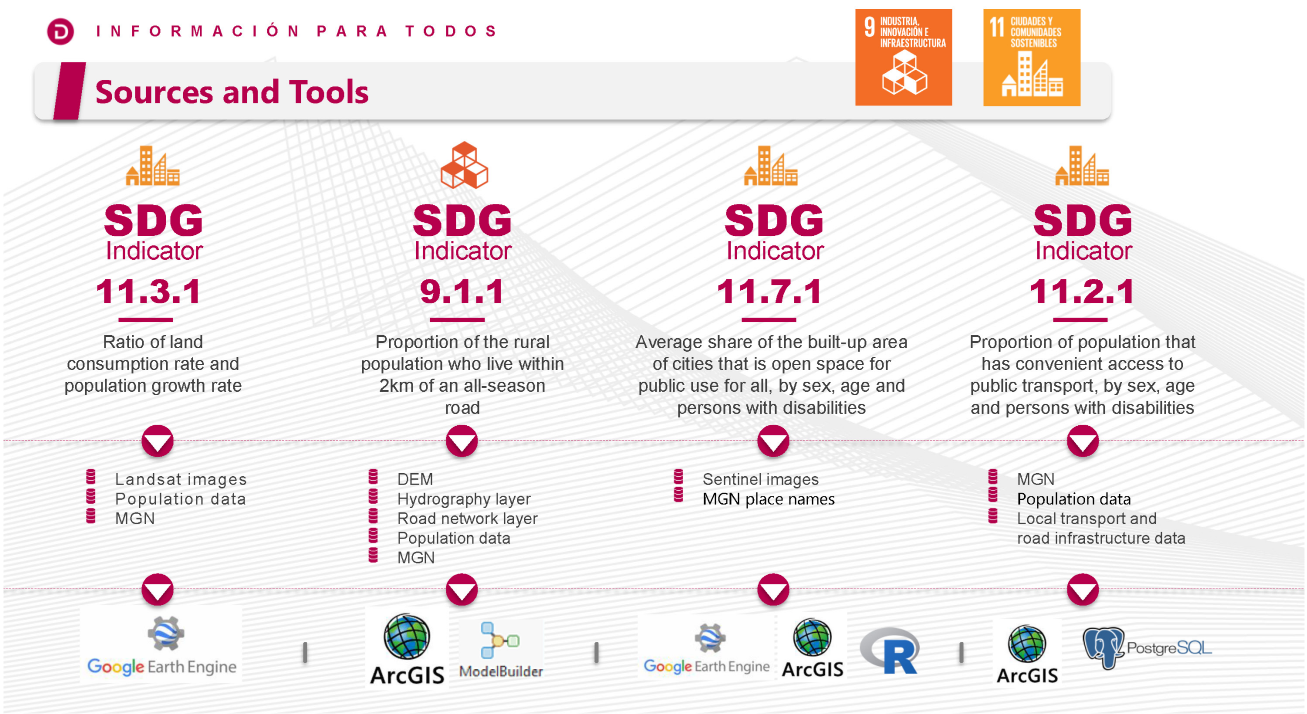


Figure 3-1: Colombia's use of geospatial data for SDGs monitoring (DANE, 2018)

In Ghana, the Ghana Statistical Service (GSS) has partnered with Vodafone Ghana and the Flowminder Foundation to produce official statistics from mobile phone data (using call detail records, or "CDRs"). The project relies on the use of de-identified (anonymized) and aggregated telecommunications data provided by Vodafone Ghana to generate population predictions. When combined with traditional data sources, such as household surveys, this provides useful information on the mobility and characteristics of the population, which can be used for a wide range of humanitarian and development applications. The GSS is also combining remote sensing data with ground-truthing from citizen-generated data to support the monitoring of marine litter (14.1.b).

The potential for using geospatial and EO data for SDGs monitoring is also being tapped by many countries with existing geospatial capabilities and programs. For instance, Japan's Space Agency (JAXA) is using the launch of new Japanese satellites (Himawari-8 and-9) to estimate surface aerosol concentrations for monitoring indicator 11.6.1, with methodologies also under development for other SDGs' indicators including the ratio of land consumption (11.3.1), open space in cities (11.7.1), and the mountain green cover index (15.4.2). In China, the Chinese Academy of Science Big Earth Data Engineering Program (CASEarth) has studied the contribution of EO data for monitoring a range of indicators across six SDGs (2, 6, 11, 13, 14 and 15) (Chinese Academy of Science, 2020). The National Statistics Office of Mongolia is using ArcGIS to undertake experimental estimates for SDGs indicators on road access (9.1.1) and electricity access (7.1.1). And Malaysia's Department of Statistics is collaborating with the Malaysian Space Agency to estimate the proportion of the population living in remote and inaccessible areas (17.19.2), and has partnered with the Malaysian Rubber Board through the ONS-UNECE Group for Machine Learning project to use satellite imagery and machine learning to estimate rubber plantation areas (2.4.1). In Fiji, the Fiji Bureau of Statistics is using QGIS open software and medium resolution Sentinel satellite data to provide experimental land cover accounts and estimate indicators relating to forest cover (15.1.1), urban areas (11.3.1), and agricultural land-use change.

The potential for EO applications to support SDGs monitoring is also being advanced through a range of global initiatives, including Open Data Cube (**Box 3-3**). The National Institute of Statistics and Geography (INEGI) in Mexico has recently leveraged the Open Data Cube to expand their geospatial capabilities, including new SDGs monitoring applications such as water-related ecosystems (6.6.1) (Luis et al., 2019). The Digital Earth Africa initiative has also utilized the Open Data Cube platform to provide access to data, analytical tools, and training for countries in the region (Box 3-3).

Box 3-3

Open Data Cube and Digital Earth Africa

The Open Data Cube (ODC) is a non-profit, open-source project that was motivated by the need to better manage and utilize satellite data and was originally developed by Geoscience Australia. The Committee on Earth Observation Satellites (CEOS) is a founding partner in the ODC initiative which seeks to provide a data architecture solution that has value to its global users and increases the impact of EO satellite data. The initiative aims to minimize the time and specialized knowledge required to access and prepare satellite data and provide an open and freely accessible exploitation tool to support its community to realize new applications. At its core, the ODC is a set of Python libraries and a PostgreSQL database that helps users work with geospatial raster data.

A flagship application of the ODC is Digital Earth Africa (DE Africa), providing not only data but also analytical tools, maps and online training. The platform has been used to look at issues of flooding, coastal erosion, agriculture, and urbanization across the continent. The establishment of DE Africa has been funded by US-based Leona M. and Harry B. Helmsley Charitable Trust and the Australian Government. Digital Earth Africa will support all countries in the region to drive progress towards national development priorities and the SDGs, in particular those relating to zero hunger (SDG 2), clean water and sanitation (SDG 6), industry, innovation and infrastructure (SDG 9), sustainable cities and communities (SDG 11), climate action (SDG 13), life below water (SDG 14), life on land (SDG 15).

3.2 Emerging Innovations from the Research Community in Big Data for SDGs Monitoring

In the research community, innovations in the use of big data sources to support SDG monitoring are rapidly progressing, with many of the latest advancements published in the academic literature. A systematic review of the academic literature since 2015 was undertaken as an input to this paper. The study used a query protocol to identify **100 recent published datasets derived from big data** and collated information on these datasets (Allen et al., 2021). A repository of the innovations is also available online (www.bigdatasdgs.com), mapped to each of the SDGs.

The results from the review highlight recent advancements by the scientific and research community in cooperation with a range of partners from government, international organizations, the private sector, and civil society to harness big data to support the monitoring of SDG indicators. **Overall, considerable coverage of the SDGs is evident in the 100 publications and datasets reviewed, which corresponded to 15 goals, 51 targets, and 69 official indicators**. This also highlights significant thematic coverage in the datasets produced, which include economic, social, and environmental issues. In some instances, **there is also evidence that these advancements have been adopted by international organizations and governments to support global and national monitoring.**

Figure 3-2 provides a snapshot of some of the key results from the analysis, including the proportion of papers reviewed that corresponded to each goal (top line, colored circles), the number of papers addressing specific official SDG indicators (green shaded rectangles with indicator numbers), and the share of the total number of official indicators for each goal that were addressed in the papers reviewed (brown shaded diamonds with percentages, bottom line). Taken together, these attributes highlight thematic areas with greater activity in terms of research on big data for monitoring SDGs. **Overall, the largest share of papers corresponded to SDG 15 on life on land (21%), SDG 6 on clean water and sanitation (15%), SDG 1 on poverty (12%), and SDG 11 on sustainable cities (12%)**. There was also considerably stronger coverage of official indicators for SDG 3 on health (75% of SDG indicators), SDG 15 (64%), and SDG 6 (64%). This suggests that these goals are likely the dominant thematic areas for recent research using big data to support SDG monitoring, with many potential use cases.

SDG indicators with particularly strong coverage in the papers reviewed (**Figure 3-2**) included 1.1.1 and 1.2.1 (9 papers) relating to income poverty, indicator 6.6.1 (9 papers) on the extent of water-related ecosystems, 15.3.1 (8 papers) on the area of degraded land, and 15.1.1 (7 papers) on forest area. Other indicators with comparatively higher coverage of five or six separate studies included 2.4.1 (sustainable agriculture), 6.4.1 (water use efficiency), and 11.3.1 (land consumption rate). Of the indicators listed with a large number of multiple studies, only 2.4.1 and 11.3.1 are classified as Tier 2 indicators (**Figure 3-2**, black boxes).

The different innovations and use cases for big data sources and SDGs monitoring are explored in more detail for each of the goals Appendix 1.

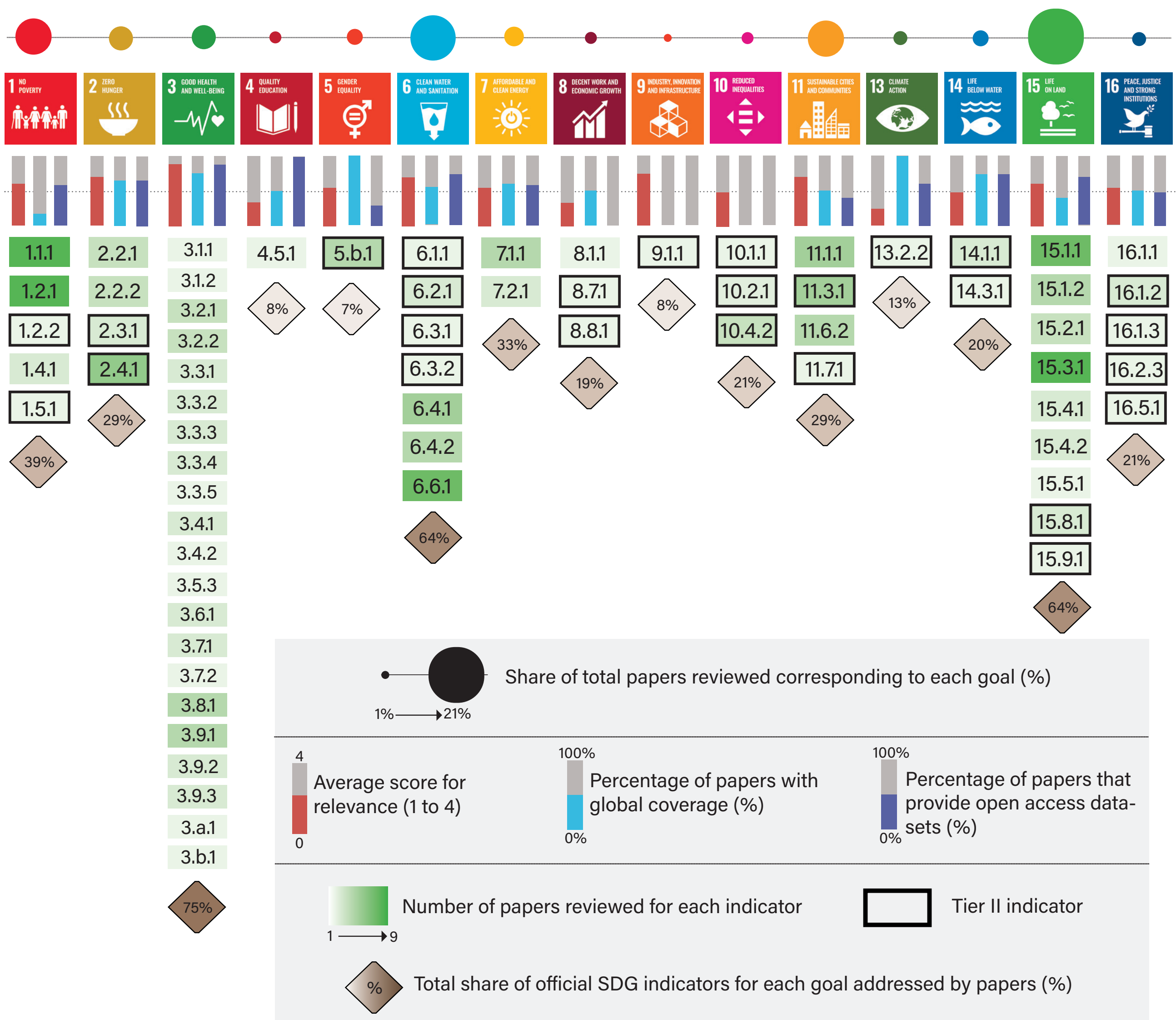


Figure 3-2. Coverage of goals and indicators of the SDGs in the papers reviewed and other key attributes (Allen et al., 2021) Refer to inset legend for interpretation.

The results from the systematic review demonstrate a range of potential use cases for different types and sources of big data. **Overall, the dominant big data types** included satellite or EO data (82 papers), other sensor network data (12 papers), tracking data (13 papers), administrative data (11 papers) and opinion or behavioral data (9 papers). None of the papers reviewed used transactional data. Approximately 25% of papers also reported using survey data in their analysis, more commonly for socio-economic goals and indicators.

Figure 3-3(a) links the big data types to different SDGs, where the size and shading of the outer spheres reflect the number of new datasets that used each type. The dominance of satellite and EO data is evident and was used for all SDGs (except SDG 5), with the greatest number of applications for SDGs 1, 6, 11, and 15. Other sensor network data (including from road and climate sensors) were used more commonly for SDGs 2, 3, 11, and 13. Tracking data (primarily from mobile phones) were used more commonly for SDGs 1 and 7, while opinion data (including social media data from Facebook and Twitter) were used more commonly for SDGs 1 and 5. Administrative and survey data were primarily used for social indicators relating to SDGs 1, 2, 3, and 4.

Within these big data types, **emerging innovations also used a range of different data sources (Figure 3-3(b))**. Sources of multispectral satellite and EO data were the most reported, including Landsat (23 studies), MODIS (15 studies), Sentinel (12 studies), and QuickBird (2 studies). These data sources provide predominantly moderate to high-resolution imagery and land surface reflectance datasets that are more commonly used for environment-related goals, including SDGs 6, 11 and 15. Other common data sources included night-time lights (NTL) imagery data from the Visible Infrared Imaging Radiometer Suite (VIIRS) (18 studies) and OpenStreetMap (OSM) crowdsourced data on roads and infrastructure (10 studies), particularly for SDGs 1, 7, and 11. Dominant sources for population data included WorldPop (6 studies) and Gridded Population of the World (GPW) (3 studies), while the Global Human Settlement Layer (GHSL) was also used for defining urban areas (7 studies). Social media data from Facebook and Twitter was only used for SDGs 1, 5 and 16, while call detail records (CDRs) were used for SDGs 1 and 3. Overall, studies on SDG 1 reported the greatest variety of data sources, which highlights greater experimentation with different data sources for reporting on poverty compared to other goals. In terms of the versatility of data sources, NTL was used in studies corresponding to the largest variety of goals (8 SDGs), followed by MODIS (7 SDGs) and Sentinel (7 SDGs). A brief description of these main big data sources is provided in **Box 3-4**.

The dominance of EO and satellite data as a primary source for new datasets for monitoring the SDGs is not surprising, given their accessibility and comprehensive geographic coverage. Reviews of the use of big data sources by countries in their national SDGs reporting (Voluntary National Reviews, VNRs) have also found a dominance in the use of EO and geospatial data sources (UNESCAP, 2021). One of the likely reasons for this is the field of remote sensing has been operational since the 1970s and therefore is more accepted and mature globally. For social media and mobile data, applications remain largely in the research and pilot phases.

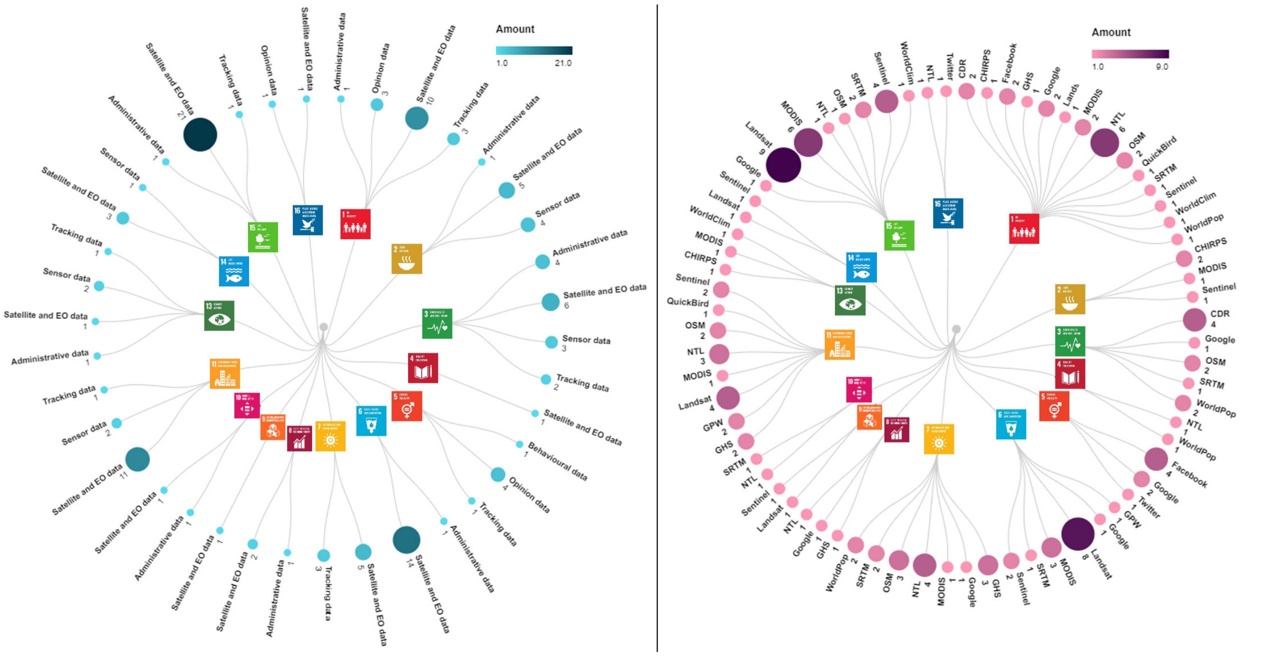


Figure 3-3 Links between the SDGs and the common big data types (a, left) and sources (b, right) reported in recent research studies. (See Box 3-9 for interpretation of sources)

Box 3-4

Common Sources of Big Data Used in Recent Studies Using Big Data Sources for SDGs Monitoring

Some of the common sources of big data used in recent studies include the following (see also Figure 3-3):

- **CDRs** – call detail records provide information about calls made over a phone service. The record is generated for billing purposes and contains various attributes including the call time, duration, completion status, source number, and destination number.

- **CHIRPS** – Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) is a 35+ year quasi-global rainfall data set. The dataset uses “smart” interpolation techniques and high-resolution, long period of record precipitation estimates based on infrared Cold Cloud Duration observations.

- **GHSL** – Global Human Settlement Layer is an open dataset and provides global spatial information about the human presence and settlement on the planet over time. This is in the form of built-up maps, population density maps, and settlement maps. The framework uses heterogeneous data including global archives of fine-scale satellite imagery, census data, and volunteered geographic information. The data is processed fully automatically and generates analytics and knowledge reporting objectively and systematically about the presence of population and built-up infrastructures.

- **Landsat** – NASA’s Landsat program offers the longest continuous global record of satellite imagery of the Earth’s surface launched in 1975. Currently, both Landsat 7 and Landsat 8 are in a near-polar orbit so that each location on Earth is measured by one or the other every eight days. As the Landsat satellites orbit, the instruments capture scenes across a swath of the planet that is 185 km (115 miles) wide. Each pixel in these images is 30 m across.

- **MODIS** – NASA’s Moderate Resolution Imaging Spectroradiometer has a much wider reviewing swath (2,330 km wide) and sees every point on the plate every 1-2 days. MODIS is ideal for monitoring large-scale changes in the biosphere.

- **NTL** – Night-time light data from the US Defense Meteorological Satellite Program (DMSP)/Operational Linescan System (OLS) and the Visible Infrared Imaging Radiometer Suite (VIIRS) on the Suomi National Polar-orbiting Partnership satellite provide satellite imagery of the Earth’s surface at night. Since the 1990s, remotely sensed NTL imagery has been shown to correlate with socioeconomic parameters including urbanization, economic activity, and population.

- **OSM** – OpenStreetMap is an open crowdsourced global map of the world. Map data is collected from scratch by volunteers performing systematic ground surveys using tools such as a handheld GPS unit. The availability of aerial photography and other data from commercial and government sources has added important sources of data for manual editing and automated imports. It is published under an open license that allows anyone to access, use, and share the data.

- **Sentinel** – The Sentinels are the satellites of European Space Agency (ESA), designed to deliver a vast amount of data and imagery for Europe’s Copernicus program. Sentinel 1 and 2 deliver day-and-night radar imaging and multispectral high-resolution imaging for land and ocean monitoring.

- **SRTM** – NASA’s Shuttle Radar Topography Mission provides high-resolution topographic data and provides digital elevation models on a near-global-scale.

- **WorldPop** – provides high-resolution, open and contemporary data on human population distributions at 100m spatial resolution, allowing accurate measurement of local population distributions, compositions, characteristics, growth, and dynamics.

- **WorldClim** – is a database of high spatial resolution (1 km) global weather and climate data commonly used for spatial modelling.

With regard to the case of geospatial data, **the rapid pace of recent advancements in SDG applications has also been largely driven by ubiquitous and free public access to satellite imagery as well as cloud computing infrastructure and algorithms**. In interviews with several expert authors involved in big data research, the Google Earth Engine (GEE) platform was considered the most significant innovation in the last decade. The GEE platform incorporates free optical and radar imagery from Sentinel 1 and 2 from the European Space Agency and is free to use for researchers and not-for-profit purposes. Datasets from the National Oceanic and Atmospheric Administration (NOAA) and National Aeronautics and Space Administration (NASA) such as a processed NTL product, are also readily available, and the cooperation of space agencies has been critical in providing public access to new data. While some research institutions and NSOs have high-performance computing clusters capable of undertaking large-scale analyses at a very high resolution, they are often in demand and unavailable for the long time frame needed for these analyses. A key incentive for Google in providing this infrastructure is reportedly related to social responsibility, however there is also scope for commercial applications and a new income stream. For some projects, such as the World Settlement Footprint (Marconcini et al., 2020), the GEE staff have played an active role in optimizing the analysis (**Box 3-5**). For others, they simply provide the infrastructure and are not a formal partner. Data used for analyses in GEE also must be deposited on the Google servers, which may raise privacy concerns depending on the nature of the data. Alternatives to GEE were also mentioned, including Amazon's AWS cloud computing services, Microsoft's planetary computer, and the European Data and Information Access Services (DIAS).

Box 3-5

World Settlement Footprint (WSF) (Marconcini et al., 2020)

Advancements in global datasets produced from big data that have relevance for SDGs monitoring are being used widely by international organizations. For example, researchers from the European Space Agency and the German Aerospace Centre have identified a set of "Champion Users" to test and provide feedback on their new data products, building on a research partnership with Google, that produced the World Settlement Footprint (Marconcini et al., 2020). New products include World Settlement Trends (1985-2015), a higher-resolution 2019 WSF layer (10m resolution), coupling of the WSF with population, income, and CO2 emissions data layers, new layers on impervious surfaces and green surfaces, and building heights. These new products have many potential applications relevant to the SDGs and have gained considerable interest from Champion users, which include the Cities Resilience Group at the World Bank, disaster risk groups in UNHABITAT and other UN organizations, and the Red Cross. The main applications include disaster risk assessment (e.g., floods, landslides, and earthquakes), monitoring of urban form (e.g., change in urban development and densification of cities), and monitoring of displaced persons and refugee camps. This can help inform planning decisions and resource allocation decisions, as well as contribute to monitoring SDGs indicators such as on land-use efficiency (11.3.1).

3.3 The Challenge of Ensuring Big Data Innovations Benefits the National Statistical System and SDGs Monitoring

The datasets derived from big data reviewed above and in Appendix 1 could serve a range of different monitoring objectives. For example, some studies provided **new global datasets** for monitoring official SDG indicators, such as for 6.6.1 (Pekel et al., 2016), 11.3.1 (Schiavina et al., 2019), 15.3.1 (Meyer et al., 2020, Hengl et al., 2017) and 15.4.2 (Bian et al., 2020). While other studies aimed to develop and test novel methods for **more timely and disaggregated estimates** to fill gaps in indicators such as on poverty (Yeh et al., 2020) or electricity access (Falchetta et al., 2019). A number of studies also provided **new complementary datasets** to add more depth to official indicators including for monitoring essential biodiversity and climate variables (Hansen et al., 2019, Funk et al., 2015) or multidimensional poverty (Pokhriyal and Jacques, 2017) as well as globally consistent **subnational gridded estimates** for SDG indicators that are more useful for local-scale monitoring and decision making, particularly with regard to health indicators (GBD 2019 Diseases and Injuries Collaborators, 2020, GBD 2019 Universal Health Coverage Collaborators, 2020, Osgood-Zimmerman et al., 2018, Graetz et al., 2018).

Within the follow-up and review framework of the SDGs (UNGA, 2015), national governments and NSOs are primarily responsible for data collection, coordination, validation and quality assurance (United Nations General Assembly, 2017, UN Statistics Division, 2017). And while academic and research institutions are not official partners, they play an important role in developing metrics and methods and can serve as analytical partners, knowledge brokers, convenors, and centers of excellence (Sachs et al., 2020, Kickbusch and Hanefeld, 2017). **With many new datasets emerging from experts at the forefront of big data research, there is a growing need to clarify how best to t these new global datasets and methods to national monitoring and reporting processes.** From the datasets reviewed, it is evident that a broad range of partners beyond academia are engaged in global research efforts, in particular international organizations, space agencies, the private sector, and civil society organizations. However, there are few examples where NSOs are involved, and **it is often unclear how new global datasets could or should be used by national governments to support or complement their reporting.**

There are some exceptions, but these examples are few. For example Giuliani et al. (2020) develop a flexible and scalable approach for countries to generate their own statistics for monitoring indicator 15.3.1 on land degradation using open-source software and either free global datasets or their own national datasets (**Box 3-6**). The World Bank's Light Every Night open data repository is another recent example (**Box 3-7**), which builds on advanced methods developed by researchers (Falchetta et al., 2019) and provides open access to standardized and analysis-ready geospatial (NTL) data along with code, tools and training for countries to produce timely estimates for indicator 7.1.1. Additionally, the Freshwater Ecosystems Explorer for indicator 6.6.1 is now being tested by countries to support national monitoring (**Box 3-8**).

Box 3-6

Mapping Land Degradation to Support National SDGs Monitoring

An example where international organizations have engaged with countries to enable the use of geospatial data for SDGs monitoring was highlighted for land degradation (15.3.1) (Giuliani et al., 2020). The custodian agency (UNCCD) led the process with the objective being to help countries to set a baseline for 2015 as well as the 2030 targets. A scientific framework for reporting on the indicator (Cowie et al., 2018, Orr et al., 2017) was developed through academics on the Science-Policy Interface of UNCCD and technical partners from NASA and the European Commission Joint Research Center (EC-JRC) assisted with developing an open-source tool that countries could use to produce the indicator as well as with training and capacity-building. The project was financed through the Global Environment Facility (GEF), and countries were incentivized to join through the availability of follow-up funding once targets were set. This holistic design, bringing in a range of partners to support country-led reporting provides a good case study of bridging the research-policy divide.

Box 3-7

Monitoring of Electricity Access Using Satellite Imagery: The World Bank's Light Every Night Portal

By leveraging recent advancements in the use of night-time lights (NTL) and population datasets combined with machine learning algorithms to estimate electricity access, the World Bank has launched the Light Every Night initiative to support national SDG monitoring. Light Every Night is a comprehensive data repository of NTL satellite imagery collected from two sensors over the last three decades: the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) with data from 1992-2017, and the Visible Infrared Imaging Radiometer Suite (VIIRS) Day-Night Band (DNB) with data spanning 2012-2020.

The DMSP-OLS and VIIRS-DNB sensors capture various sources of low-light emissions from Earth. These include sources that indicate aspects of human activity, such as city lights, gas flares, fishing boats, and agricultural fires, while also capturing other night-time lights phenomena (e.g., auroras).

The World Bank worked in collaboration with the National Oceanic and Atmospheric Administration (NOAA) and the University of Michigan to publish this repository, and it was designed from the ground up to be analysis-ready. The underlying data are sourced from the NOAA National Centers for Environmental Information archive. Additional processing by the University of Michigan enables access in Cloud Optimized GeoTIFF format and search using the Spatial Temporal Asset Catalog (STAC) standard. These standards are part of the growing Analysis-Ready Data ecosystem that is improving access to geospatial data sets and enabling broader audiences to readily discover, process, and analyze geospatial data.

The datasets are published along with tutorials and training to support countries that wish to explore and use the data to produce estimates for access to electricity. For further information see the **World Bank's Light Every Night** portal and **tutorials**.

Box 3-8

Cross-Sector Collaboration for Global Monitoring of Surface Water

In 2018, the UN Environment Programme (UNEP) was made the custodian of 26 key SDG indicators, and the organization launched a special unit on environment statistics to meet this sizable obligation. In particular, UNEP was tasked with monitoring indicator 6.6.1 on the change in extent of water-related ecosystems over time, a broad topic that UNEP separated into several sub-indicators. However, a survey of NSOs found that only approximately 40 countries had data relevant to these water ecosystem topics, and available data was not necessarily comparable. While continuing to work with countries on developing statistical capacity, UNEP recognized the need for new global data products that could help fill these knowledge gaps and formed a partnership with Google to create a global indicator of surface water that has since been incorporated into official SDG reporting.

The partnership sprang from an existing collaboration between the Google Earth Outreach team and the European Commission Joint Research Center (EC-JRC) who had worked together to create the Global Surface Water Explorer. This tool analyzed Landsat data using Google's computing infrastructure to map surface water around the world at a 30m resolution. UNEP initially contacted Google to express interest in using the existing Global Surface Water Explorer for SDG monitoring. Shortly thereafter, representatives from UNEP and Google held high-level discussions on the sidelines of the UN High-level Political Forum on Sustainable Development (HLPF) in July 2018 and created a detailed work plan and signed a memorandum of understanding (MoU). Although Google provided the data free of charge, and the MoU was non-binding, forming an agreement was important for building and maintaining trust in the partnership. UNEP and Google then collaborated on a public user interface for the surface water map, which they previewed at the UN World Data Forum in October 2018 and launched at the UN Environment Assembly in March 2019. Data from the partnership is now featured in the Global SDG Database, making it the first cross-sector partnership of this kind to inform the SDGs. Furthermore, UNEP has received several inquiries from Member countries about the data, and the surface water map is used to inform policymaking.

As of 2019, the freshwater data had been tested by the Namibia Statistics Agency and Statistics Canada, and the United Kingdom was helping with associated pilot applications in Kenya and South Sudan. For more information, visit **www.sdg661.app**.

Greater use of such scalable applications and tools would be helpful to bridge the divide between global research and new datasets derived from big data and enable countries to harness emerging methods and datasets for national reporting. There is considerable opportunity for many of the regional or global datasets reviewed to be produced and disseminated in a similar standardized format to form part of a global "analysis-ready data ecosystem." These data products based on global modelling can help to fill data gaps and ensure there are some data available for all countries (Campbell et al., 2020). The Marketplace developed by the UN Global Working Group on Big Data could provide one platform for consolidating, standardizing, and distributing such non-traditional global datasets derived from big data for easy access by countries, along with guidance on their applications for SDG monitoring. **However, while providing open-source global datasets is useful, issues will remain around interpretation, verification, understanding and ultimately the use of these datasets by national stakeholders.** Experience from more mature global data initiatives such as the GBD (**Box 3-9**) underscores the importance of ensuring that global datasets are developed alongside national capabilities and information systems.

Box 3-9

Global Burden of Disease Research Collaboration

A leading example of the benefits of broad collaborative partnerships is evident in the Global Burden of Disease (GBD) studies reviewed, which include thousands of collaborators, provide gridded datasets covering 37 SDG health-related indicators across several goals, and are underpinned by an ongoing global research partnership, led by the University of Washington. Annual results are reviewed and published in cooperation with The Lancet journal, and model code and datasets are made available through the Global Health Data Exchange along with visualization through other online tools. This collaboration provides a useful model that could be expanded to other thematic areas to better leverage research and innovation in support of global SDG monitoring. However, while these datasets are widely seen as a global public good, concerns have been raised that the approach lacks transparency and may hamper the development of effective national health information systems (Shiffman and Shawar, 2020). Increased accountability and pluralism (including greater involvement of low-income countries) is needed, and that both the global health metrics and national health information systems agendas advance synergistically.

Interviews with experts involved in big data research also highlighted that integrating new datasets to NSO processes for use by national policy makers was considered the "last mile to go." This engagement was considered particularly important to move beyond pilot projects and develop production pipelines with sustainable financing. However, a number of barriers to this were cited, including gaps in technical skills, limited opportunities for engagement between specialized experts and NSOs, and methodological challenges.

In terms of methodology, **datasets derived from big data are a proxy measure or estimate that provides a certain likelihood or confidence level that a phenomenon occurs.** Innovative sources of data are therefore not perfect substitutes for traditional data sources on which many official statistics are based and are often collected as a by-product rather than designed by statisticians for a particular purpose. As such, **they do not have the same quality as well-designed survey data and the current standards and conventional statistical methods do not apply.** While official statistics are based on scientifically rigorous guidelines vetted by experts to ensure representativeness, accuracy and reliability, there are also concerns about potential bias and representativeness of big data sources. For example, concerns have been raised regarding the accuracy and the need for national ground-truthing and validation for estimates derived from the commonly used NTL dataset (Andersson et al., 2019) or mobile phone data (Blumenstock, 2018). **It is important that estimation methods that incorporate big data do not aim to replace conventional methods, and the integration of conventional and innovative sources can help to ensure accuracy and reliability while improving granularity, coverage, and timeliness.**

With regard to engagement with NSOs, the use of global geospatial datasets for national monitoring provides a useful case study. Efforts in areas such as land cover and land degradation monitoring have largely involved international custodian agencies, international conservation organizations, space agencies, and research institutes with funding from global donors. Engagement at the national level has been with line ministries and focal points (e.g., for environmental conventions) rather than NSOs. One reason for this is the strict standards applied by NSOs and the slow uptake of new data sources and methods. While there is a preference for using national datasets for reporting, global datasets have been substituted where national datasets are lacking. However, **the validation of these global datasets can be challenging for governments due to the opaque disclosure of assumptions, calibration, and verification procedures which may not adequately consider local context.** And for decision makers who are trying to set targets or allocate resources, there may be agreement that some data is better than none even if it is less accurate, however this may not be a widely-held opinion of NSOs or other stakeholders.

While many existing EO projects can contribute to SDGs monitoring, there is often a missing link between national mapping or geoscience agencies and NSOs. As noted, countries such as Mexico have aimed to address this by merging these capabilities in their National Institute of Statistics and Geography. Although some countries may have sophisticated geospatial and data science capabilities, including via national space agencies and geoscience agencies, discussions with experts suggest that the technical methods and infrastructure requirements for deriving new datasets from EO and geospatial data may be prohibitive for many countries. Fortunately, there are a handful of global initiatives under way to provide tools and guidance for using EO data to benefit SDGs reporting and support decision-making (**Box 3-10**, and **Boxes 3-6** to **3-9**). Such approaches are useful to ensure the dual objectives of filling gaps in global monitoring on the SDGs and building national capabilities to produce and validate their own datasets are met.

Given the broad utility of EO data and other big data sources for SDGs monitoring, building awareness, and developing maturity in skills and capabilities more widely in NSOs will be critical to mobilizing opportunities. These capabilities clearly exist in a range of partners globally. From the systematic literature review, key collaborators include international and regional organizations (the UN, development banks, GEO, and the European Commission), space agencies (USA, UK, German, and European), national government agencies (NSOs, scientific and geological survey agencies, line ministries), research institutes (IIASA, EC-JRC), non-government organizations and think tanks (International Union for the Conservation of Nature, The Nature Conservancy, World Resources Institute, Flowminder Foundation, MindEarth), the private sector (Google, Orange Telecom, EOMAP), and donors (philanthropic foundations, GEF), among others. **These partners play a range of different roles in the big data ecosystem and will be critical in enabling NSOs to harness the potential of big data for SDG monitoring.**

Box 3-10

Earth Observation Toolkit for Sustainable Cities and Human Settlements

Earth Observations (EO) data acquired remotely by space-borne, airborne, and in situ sensors, and ground-based observations can provide reliable, timely, and continuous sources of information for SDG 11 and related urban issues. However, many national and local level governments and organizations are unfamiliar with or lack the skills and capacity required to apply Earth Observations in support of urban policy, planning, monitoring, reporting, and operational decision-making.

The Earth Observations Toolkit for Sustainable Cities and Communities was developed to support local communities, cities, and countries in understanding the value and usefulness of Earth Observations for SDG 11 and the New Urban Agenda. Key toolkit components include guidance on and links to relevant EO datasets, tools that can (and already do) support the definition and refinement of SDG 11 indicators, as well as innovative approaches and national or city-level experiences of EO uses for SDG 11 and the New Urban Agenda. The toolkit also aims to facilitate engagement among local communities, cities, national agencies, and EO experts, and to promote knowledge-sharing and collaboration between cities and countries.

4. Partnerships for Harnessing Big Data for SDG Monitoring

The core value proposition of NSOs is providing trusted and quality statistics and their operational model is based on a data value chain from collection through processing to dissemination and use (Open Data Watch, 2018). Recent examples of global and national applications of big data sources to support SDG monitoring highlight a complex ecosystem involving multiple stakeholders and partners from governments, private enterprises, academia, international organizations, civil society, and technology service providers. **This new data ecosystem challenges the traditional model, requiring NSOs to tap new data sources, adopt new techniques, utilize new platforms and work with new partners** (Cazarez-Grageda and Zougbede, 2019). Fortunately, considerable opportunities exist for partners to support NSOs and governments with their national SDG reporting obligations, by providing the necessary data sources, skills, infrastructure, knowledge, and financing to deliver big data projects.

The increasing importance of partnerships for data and statistics is evident in a range of recent studies and guides. These studies address various aspects such as public-private partnerships (Robin et al., 2016), business models and private sector incentives (Klein and Verhulst, 2017), smart data strategies and capabilities (OECD, 2018), establishing and using partnerships in big data statistical projects (UNSD and UNECE, 2015, Partnerships Task Team, 2014), and handbooks on using mobile phone data for official statistics (UNGWG for Big Data, 2019, DIAL and Data-Pop Alliance, 2021). While they do not generally focus on monitoring of SDG indicators, they highlight several important considerations for big data partnerships, including the types of partners, incentives for partnerships, access and business models, and barriers to the uptake of big data sources.

The literature review and our interviews with experts from governments, NSOs, and big data researchers highlighted **a number of key challenges and barriers for developing big data partnerships and projects.** Overall, these vary considerably depending on the type of big data, ownership of the data (private versus public), the working environment, capabilities, and legal context. Additionally, the degree of maturity in several of these areas shapes the big data partnership arrangements and processes, and differs from country to country. As such, **there is no one-size-fits-all solution to overcome the various obstacles, and partners can play many different roles. To develop effective partnerships, it is important to understand the partnership landscape, the key barriers and challenges, and effective mitigation measures.**

4.1 Understanding the Partnership Landscape – Types, Roles and Partnership Models

Partnerships between national governments, NSOs, colleagues from academia, the private sector, international organizations and civil society provide the most promising means to harness big data for SDG monitoring. **As evidenced by recent experience and innovations, these partners engage in big data projects through a range of different roles and working modalities (Table 4-1).**

While partnerships between NSOs and data providers for accessing big data are often what comes to mind, the partnership landscape is more diverse and complex. This is due to big data's many attributes (**Table 2-1**), which require advanced and novel techniques and technologies to enable the capture, storage, distribution, and analysis of the information. For example, **in addition to data providers, statistical organizations may need to partner with technical service and technology providers** to gain access to these capabilities if they are not available in-house. To initially explore the possibilities and contributions of big data sources, **knowledge brokers or convening organizations** also may be needed to raise awareness and build trust. Additional funding is often needed to get big data projects off the ground, which can benefit from an **executive sponsor to champion** a business case for internal funding allocations, or an **external donor** willing to provide funding. Finally, a **legal or data privacy partner** to assist in navigating the regulatory landscape is another important and often overlooked partner.

Table 4-1 Common Types of Partners Involved in Big Data Projects

PARTNER TYPES	ROLE	COMMON EXAMPLES
Data Users	Custodians and users of national statistics.	NSOs, Government/ policy makers, NGOs, International organizations.
Data Providers	The public and private owners of big data sources, which may be made available open access or on a cost-recovery or commercial basis.	Space agencies, mobile phone companies, social media companies, government agencies, and NGOs.
Analytical Partners	Provide the advanced skills in design and analysis (e.g., developing algorithms and programs) that are beyond the capacity of governments or NSOs. Analysis partners often provide the know-how of acquiring the underlying data, transforming raw data into the intended data structure, and applying techniques to produce the end results.	Flowminder, Google, academic institutions, research institutions, geoscience agencies, and domain experts.
Technology Partners	Provide products for storage, computation and analysis through stand-alone software, full platforms or hardware. This includes cloud-based solutions from private partners that are delivered and priced in different ways. Technology partners also often combine solutions with training and consulting services to facilitate analysis.	Google, Microsoft, Amazon, Esri, and cloud computing providers.
Knowledge Brokers/ Convenors	Convene and coordinate different actors, build relationships between partners, and raise awareness of use cases.	GPSDD, TReNDS, PARIS21, CEPEI, P4R, and International Organizations.
Donors/Funders	Provide financing for big data projects.	Philanthropic foundations, NGOs, international organizations, and finance ministries.
Champions/ Sponsors	Build awareness, alliances, and facilitate funding.	Political leaders, government executives, and private sector leaders.
Regulators	Regulate data use, enforce data protection laws, and approve big data projects.	Privacy commissions and government regulators (e.g., telecoms).
Legal/Privacy Partners	Provide legal advice on data use, privacy, contractual arrangements, and data sharing agreements.	Government privacy officers, and Legal units.

The capability of NSOs to harness big data and the different partnership modalities that are appropriate are largely determined by the governance and technical context and maturity in different countries. Due to these differences, a range of different models have emerged that adapt to the strengths and weaknesses in the operating environment.

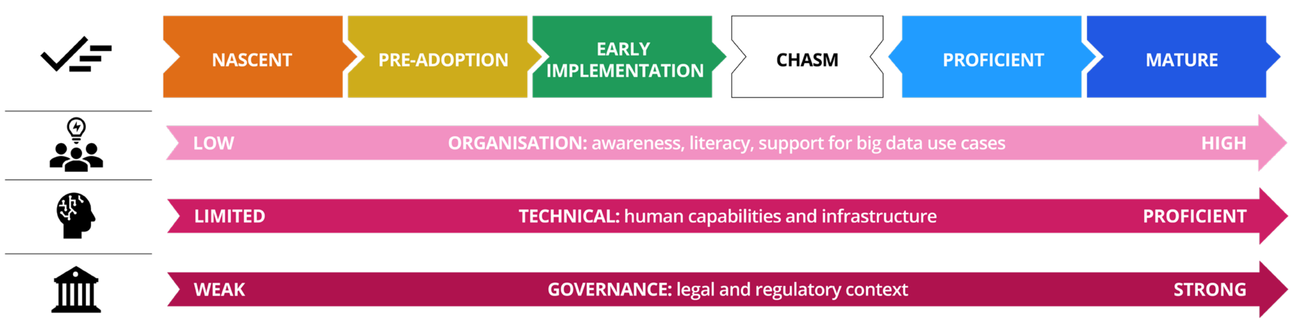
Several existing models for data access and sharing are available, including the in-house production of statistics by the data provider, the transfer of raw big data to the NSO for processing, the transfer of raw data to a trusted third-party for processing, or remote access for NSOs or third parties (Klein and Verhulst, 2017, Robin et al., 2016, UNGWG for Big Data, 2019). However, the central focus of these models is on the allocation of responsibility for data processing tasks, and there is an assumption that the use case is well understood, that a sound business case has been developed and supported, and that other initial barriers have been overcome.

Discussions with NSOs and experts also shed light more broadly on barriers to the use of big data sources for SDGs monitoring or national official statistics. These can be generally defined as a lack of awareness and difficulties in initiating projects, legal and privacy issues, and technical capability and methodological challenges. The emphasis placed on these different barriers and the types of partnerships used to overcome them varied across discussions with different countries and stakeholders. This demonstrates that most governments and NSOs are at different stages in terms of their maturity in deriving value from big data, which influences the type of partnerships that are useful for harnessing big data.

4.2 A Big Data Maturity and Partnership Model for NSOs

Some NSOs have yet to commence their big data journey, while others are quite advanced. **Maturity models** can provide an intuitive framework to understand the current state and progress of an organization and the steps needed to move to the next stage(s) of maturity. They generally include a sequence of levels - **stages** - that define a path from the lowest (initial) to the highest (ultimate) state of maturity based on a set of attributes exhibited by the NSO, usually described as **dimensions**. Maturity models have been devised for NSOs for the production of data to inform SDG indicators (Marcovecchio et al., 2018), for integrated national data systems (World Bank, 20201), as well as for big data maturity in organizations (Halper and Krishnan, 2013).

There are three dimensions used to organize the NSO's attributes (represented by the pink arrows in **Figure 4-1**: i) **Organizational characteristics**: big data awareness, literacy, and support for big data use cases; ii) **Technical capabilities**: both human and infrastructure; and iii) **Governance**: the legal and regulatory context.



	NASCENT	PRE-ADOPTION	EARLY IMPLEMENTATION	CHASM	PROFICIENT	MATURE
Data providers		● ●	● ● ●		● ● ●	● ● ●
Analytic partners		● ●	● ● ●		●	●
Technology partners			● ●		● ●	● ●
Brokers/Convenors	● ● ●	● ● ●	● ●			
Donors		● ●	● ●			
Legal partners		● ● ●	● ●		●	●
Champions/Sponsors		● ● ●	● ● ●		●	●

● ● ● critical or lead ● ● supportive ● engaged

Figure 4-1 A Big Data Maturity and Partnership Model for NSOs

For the **organization** dimension, awareness varies regarding the potential application of big data within government and the NSOs, as well as a lack of understanding of its strengths, weaknesses, and limitations. When big data is a novelty, NSOs must engage with users and experts to understand the data and potential use cases. In countries experiencing many data gaps, big data provides an alternative source, but because it is not necessarily comparable to traditional sources, it can take time for users to see the utility of these products before they are demonstrated and there may also be concerns relating to completeness, bias, and verification. In these circumstances, an important role of the NSO as the head of the national statistical system is to build confidence in the potential for new data sources, to familiarize users with potential use cases, and to bring all stakeholders on-board.

NSOs also need to better understand the limitations of different data sources, their potential biases, and methods for mitigating these. In the context of SDG monitoring, governments and users may require demonstrable proof of the potential for big data to improve monitoring and a clear business case establishing the benefits outweigh the costs. **Engagement with regional or global statistical communities and other partners can build awareness of potential use cases for big data, while such convening and brokering organizations can kick-start national discussions and raise interest in pilot applications.** Costs for initiating projects can be supported by international donors and organizations or by internal champions who can help build and present the business case for financing.

Processing big data also presents **technical** challenges – the second dimension of a maturity model. Here, skillsets in data science, high-performance database systems and computing equipment, and cloud computing platforms are touchpoints. NSOs and government agencies may not have the computing infrastructure and technical staff to transfer, store, process, and analyze big data. And while processing large amounts of data is generally within the capabilities of NSOs, they may not possess domain-specific knowledge. Capacity gaps exist in varying forms and degrees in both high-income and low-income countries, but may present more significant challenges in the latter. Big data solutions used by more developed NSSs may not be immediately useful or transferable to low-income countries, who may also lack adequate financial resources to engage in big data projects (UNSD and UNECE, 2015). However, **the public and private operators who already collect, store, process, and analyze big data sources are likely to have the infrastructure and know-how to produce insights from big data. These data providers, analytical partners, and technology partners can help to fill capability gaps, provide training, and build capacities to sustain longer-term projects.**

The third and final dimension - **governance** - comprises a variety of actors, standards, and norms that determine how data is collected, stored, and shared, as well as the suitability of data sharing arrangements and risk mitigation measures. In some countries, NSOs have a broad and clear mandate to collect data and produce official statistics. They are also vested with the right to access data from both public and private sectors, within the bounds of privacy and confidentiality (UNGWG for Big Data, 2019). In contrast, other NSOs may only have the right to access data from public authorities or lack an adequate legislative mandate.

Moreover, in many cases, legislative frameworks have not kept pace with technological advancements and do not extend to big data sources. There may also be shortcomings in the legal framework governing the processing, use, or transmission of data including cybersecurity, data privacy, and data protection. The absence of adequate legal safeguards may make the use of some big data sources in official statistics riskier. In such cases, **the legislative framework for NSOs may first need to be strengthened to enable them to take the lead in exploiting big data** to produce official statistics and SDG indicators. Establishing effective legislative frameworks that enable data sharing while safe-guarding privacy has proven challenging. **Governments will need to invest considerable efforts in establishing an effective authorizing framework. In lieu of this framework**, NSOs have used ad hoc arrangements to facilitate data sharing, often in the form of MoUs or more formalized agreements with data providers or other partners. This sets the ground rules for how big data sets can be combined, protected, shared, exposed, analyzed, and retained. However, negotiating these agreements **can create significant delays in big data projects and will require early advice from legal and privacy partners to navigate the regulatory landscape.**

Once the NSO's attributes under the three dimensions are determined, **big data maturity is categorized in one of five stages** (represented by the five chevrons at the top of **Figure 4-1**): nascent, pre-adoption, early implementation, proficient, and mature. As NSOs move through these stages, they require different partners and partnership models to gain greater value from their investments (highlighted in the lower rows and columns in **Figure 4-1**).

4.2.1 Nascent and Pre-Adoption Stages of Maturity

The **Nascent Stage** represents a pre–big data environment (Halper and Krishnan, 2013). In this stage, the NSO and NSS have a low awareness of big data, potential use cases, and value for supporting official statistics and SDG monitoring. While there may be pockets of people spread across government with an interest in the potential value of big data, there is no executive-level support or leadership. The organization may have in-house data analytics capabilities, but they have not yet been applied or adapted for big data applications.

In the **pre-adoption stage**, the NSO is starting to explore big data analytics, and staff may be studying methods, attending conferences and engaging with global partnerships. The organization may also be developing a business case for different applications of big data to complement or replace statistical products, with buy-in from an executive sponsor or champion. However, broad-based support from leadership is lacking. Sources of financing for big data projects may also be explored, including from international donors or government allocations. In this stage, the organization is beginning its big data journey, and the mindset is around experimentation (Halper and Krishnan, 2013). Business partners such as data providers or analytic partners, may be brought in to help identify priority use cases, and there may be skepticism around the potential for big data to inform statistics. Additionally, the organization may have begun to identify and collect some big data sources, but there is no defined data life cycle management, no data auditability, or lineage. There may be some experimentation with big data analytics in-house, or some external partners may be helping to develop an analytics strategy or performing advanced analytics. At this stage, organizations may also have a steering committee overseeing the program from a governance perspective. To address privacy and ethical requirements, initial engagement with legal and regulatory partners will be undertaken.

Experience from countries highlights the important role that partners can play in supporting NSOs to move through the nascent and pre-adoption stages. In Ghana, early conversations around big data commenced during the Ebola outbreak in West Africa, where the data provider Vodafone Foundation approached the Ministry of Health about the use of mobile CDR data to support response efforts. However, there were impediments associated with the Data Protection Act in Ghana. In 2017, during the first World Data Forum in Cape Town, Flowminder showcased their research on using CDRs for disaster management during a session. Subsequent meetings were held at the national level with Flowminder to explore the potential for supporting disaster management in Ghana, which raised awareness of potential use cases. To overcome the costs of initiating the project, financial support was secured early on from the Hewlett Foundation and Vodafone Foundation. Early discussions were also held with Ghana's Data Protection Commission, followed by lengthy negotiations on a formal data sharing agreement.

In the Philippines, global partners or convenors were also critical in catalyzing interest in big data in the early stages, particularly in terms of raising awareness of potential business cases and areas where big data might contribute to national statistics and decision making. An initial workshop convened by the GPSDD invited private sector data providers and analytical partners, including Google and Flowminder, as well as government and academia to assist with exploring potential use cases. Staff from the PSA as well as other government ministries also gained increasing exposure to big data use cases through their participation in other international forums and programs, particularly in areas such as price statistics, disaster reduction, and climate change. Human relationships and networks were considered fundamental in these early stages, including identifying a sponsor or champion with influence in government (e.g., legislator, minister, or influencer in the private sector). These efforts created a growing awareness and buy-in for the potential of big data and catalyzed the establishment of a government task force to oversee potential big data pilot projects. To address legal considerations and expedite projects, template data sharing agreements between the PSA and different stakeholders (e.g., researchers, government agencies, and the private sector) were also developed with advice from government, legal, and privacy officers.

Similarly, in Colombia, involvement with the GPSDD mobilized DANE to engage with CEPEI as a knowledge broker to explore the use of administrative and big data sources for official statistics. An initial step was to gain buy-in for the project by developing a business case and building trust between the data provider (Bogotá Chamber of Commerce), the statistical office, and the legal unit. Early buy-in was achieved from the President of the Bogotá Chamber of Commerce, who became an important sponsor for the project. Early engagement with the legal unit was also important for developing a suitable data sharing agreement.

ISTAT began exploring the potential of big data for official statistics in 2013, motivated by the demand for more timely statistics, as well as to improve cost and burden efficiencies. A key catalyst was their engagement in a number of regional agreements and collaborative projects from the European statistical community. These initially encouraged the NSO to explore potential use cases and pilots for big data in official statistics, which have since evolved into mature production phase projects that integrate big data into official statistics. ISTAT also continues to collaborate with the European statistical community on big data methodologies, regulation, and privacy considerations.

4.2.2 Early Implementation Stage

The **Early Implementation Stage** is typically characterized by one or two proof of concepts or pilots which become more established and production ready (Halper and Krishnan, 2013). There is generally at least one executive sponsor involved, however broader interest is likely growing as pilots deliver successful outcomes. This may lead to the establishment of a team to start to plan and strategize for a wider big data scope. During this stage, the organization will be collecting big data from data providers in different formats, and may be investing in infrastructure to take advantage of the growing data volumes or outsourcing this to technology partners. Data access and sharing agreements will be in place where needed to enable data sharing while safe-guarding privacy. The organization will also be building data analytics capabilities, usually starting with single data sources and more mature use cases and methods. Analytical partners may be providing technical support services, training, and advice to help complement and build in-house capacity. Measures are also being taken to support the recruitment and retainment of human resources for in-house data science capabilities.

In the Philippines, following initial buy-in, the primary challenge is identified by the PSA related to the capacity of the technical staff, followed by the infrastructure and resources to undertake big data projects. Building this capacity in-house was considered critical and was led by its statistical research and training institute, as well as a dedicated group for statistical methodology innovation. Assistance from peer countries, academia, and technology partners also helped to establish these capacities. For example, Esri provided ArcGIS software and training for the PSA, and academia are engaged through regional statistical committees to provide training. The PSA now has a number of pilot initiatives in place that are using big data for various purposes, including for SDG indicators. Highly advanced applications, such as the use of geospatial data for small area estimate poverty mapping, have been largely driven by international organizations, notably the ADB working with World Data Lab as a technical partner. Over time, these pilots have enabled the PSA to build in-house capacity in data sciences and improve infrastructure.

In Ghana, initial buy-in has led to the long-standing Data For Good partnership between the Ghana Statistical Service (GSS), Vodafone Ghana, and the Flowminder Foundation. This pilot project has enabled the production of rapid mobility estimates using anonymized and aggregated mobile phone data to support the government's disaster response interventions, including against COVID-19. Initial gaps in technical capacity were overcome by partnering with Vodafone. Through the project, Flowminder provides analytical support and is given remote access to raw data to produce aggregated insights for use by GSS. Training has been provided to GSS staff to use the data as a layer within a geospatial tool. Building on this success, training is now being provided to GSS staff to enable them to reproduce the analyses and shift to a more sustainable production model. Stemming from the project, there is also growing interest in other applications for measuring public access to healthcare, road networks, water, and educational facilities.

4.2.3 The Chasm – Scaling from Pilots to Production

Along with the stages of maturity, the literature suggests there is also a "**chasm**" as organizations overcome obstacles as they move from early adoption to the more mature production stages (Halper and Krishnan, 2013). Here, early challenges associated with raising awareness, big data literacy, and identifying use cases have been addressed through pilot applications. However, gaps may remain in terms of capability development and the governance context, including securing ongoing funding. The incentives and business models may not be adequate to move beyond pilot projects, and a big data management and governance strategy or roadmap may be needed. Moving to mature stages requires that big data projects become a planned and budgeted initiative treated on par with other statistical production and data integration programs. To cross the chasm, organizations need to ensure the right governance, data architecture, data life cycle management, and organizational structures are in place (Halper and Krishnan, 2013).

A key challenge for NSOs relates to building and maintaining capable technical staff, in-house data science capabilities, and infrastructure. Combining surveys and administrative sources with big data sources, including EO data, will lead to greater use of statistical modelling and machine learning methods within statistical offices. This marks a significant change in the culture and practices within NSOs (Reimsbach-Kounatze, 2015). Capabilities in these areas vary considerably between countries with different levels of maturity in data science and analytics as well as computational infrastructure. Surveys have highlighted that many NSOs still lack the technical capacities to work with big data (UNSD and UNECE, 2015) and it is difficult to attract technical experts to NSOs where salaries were not competitive. A recent survey undertaken in the context of COVID-19 highlights the use of geospatial information and technologies is yet to be mainstreamed, with the majority of NSOs in low- and lower-middle-income countries expressing clear needs to build analytical capacity and infrastructure in this area (UNSD and World Bank, 2020).

Where these capabilities are lacking, a common approach has been to use third-party technical providers to help fill capacity gaps. However, it was noted by the experts interviewed that outsourcing this to third parties may pose a risk to production continuity, as the know-how is outsourced to an external entity. Despite these risks, such partnerships for big data projects will remain critical to get things off the ground and, if well-designed, will help to build in-house technical capabilities and infrastructure that can sustain projects over the longer-term.

The difficulty in crossing the chasm to big data maturity is evident in many of the country case studies reviewed, most of which would still be considered pilot or proof of concept initiatives. There are fewer examples of projects that have been scaled up to full statistical production, and possibly none specifically related to the SDGs. National experience from NSOs using big data highlights that everything is still new – from gaining access and the collection of data, to analysis, production, and publication. To sustain big data projects, it is important to start to build the production processes from project inception and to set up the whole value chain, from access to publication. This requires early discussions along the entire production pipeline, involving methodologists, IT specialists, and statistical production and dissemination staff involved in various stages.

Many pilot projects and prototypes are tested but then fail to move beyond this stage as they have not been embedded in the data production cycle, however there are examples where big data sources have become successfully integrated into the data value chain (as explored in 4.2.4 and 5). While standardizing the business processes and architecture for the production pipeline will take time, there are opportunities to share processes, expertise, and infrastructure across organizations and regions.

4.2.4 Proficient and Mature Stages

Reaching the Proficient and Mature stages means an organization has well-established, ongoing big data programs that are executed as budgeted, planned initiatives and funding is secured. During these stages, innovation in data and data analysis becomes a core value of the organization. A range of infrastructure and technologies may be used in a unified way to support analytics, with security, recovery and backup, and performance management in place. The organization is also making use of many forms of big data, including real-time data, which are integrated into the full data value chain. Data sharing is collaborative and well managed through strong data governance policies. There are defined data life cycle management, auditability, and lineage frameworks in place as well. The organization typically has a center of excellence including a data science team which may train other groups in the use of analytics. Data analytics may be operationalized and automated as part of the business process, and governance arrangements will be in place including a steering committee to oversee progress.

Whereas countries, such as Italy, started with pilot projects, they have now moved to mature production phase projects. The most mature projects rely upon open access data sources, including web scraping data from enterprise websites (prices, labor statistics), accessing social media public APIs (sentiment index), and satellite imagery (land cover statistics). All of these projects have been developed in-house within ISTAT, but have benefited from collaboration across statistical organizations in the EU region. In the case of the land cover statistics, ISTAT has developed a completely automated pipeline using deep learning techniques for image processing, publicly available Sentinel 2 satellite imagery, and Microsoft's cloud computing platform. This highlights that technology partners continue to play a role, even in mature big data projects.

For Italy and other countries such as the Netherlands and the UK that have mature big data capabilities, technical capacity has been established over time and has benefited from regional cooperation with statistical organizations in the EU and associated big data programs and initiatives. The UK's Big Data Team and Data Science Campus as well as the Center for Big Data Statistics in The Netherlands are leading examples of NSOs embedding mature big data capabilities with clear work programs, dedicated staff and resources, and ongoing funding.

However, methodological challenges remain in terms of evaluating the quality of big data and understanding which methods are the most suitable, as well as around data privacy considerations and accessing private sector data. For example, big data projects that rely on the use of private sector data have not seen similar advancements, such as the use of CDR data from mobile network operators. In Europe, the statistical community is working to develop standardized machine learning methods, sharing training data sets, and internal and external validation methods. More mature governance arrangements are also being explored in the EU, which is reportedly reviewing regulations to facilitate access to raw data and methods to overcome privacy concerns, such as privacy preserving techniques. The European General Data Protection Regulation provides a blueprint to strike a balance between individual privacy and collective benefits by specifying the conditions under which individual data may be used for public good purposes.

5. Practical Guidance for NSOs on Partnership Processes for the Use of Innovative Big Data Sources and Methods for SDG Monitoring

A key lesson from interviews and discussions with experts is that countries and NSOs are at very different stages in their big data journey and have different levels of maturity in terms of their big data awareness and literacy, human capabilities and infrastructure, and governance and regulatory frameworks. During the journey from nascent maturity through proficiency, NSOs work with and benefit from a range of partners and partnership models. Based on this, this report puts forth a series of practical steps that NSOs can take to develop their big data maturity and partners who can help along the journey (**Figure 5-1**).

Figure 5-1. Practical Steps for NSOs to Leverage Partnerships for Big Data Innovation

5.1 Getting Big Data Partnerships off the Ground

For countries and NSOs that are yet to commence their big data journey, there is a general lack of big data literacy and awareness of the potential benefits, opportunities, limitations, and risks. Advice from NSOs who have moved beyond this stage highlights the critical role of **engaging with global and regional statistical conferences and data forums, and partnering with global and regional brokers or convening organizations** to address these issues. These could include partners such as the SDSN TReNDS, GPSDD, Data For Now, P4R, GIZ Data Lab, GEO, the Development Data Partnership, Open Data for Development, or regional and peer-to-peer initiatives and partnerships on big data (e.g., in the EU or UN Regional Commissions).

Such partners and initiatives have supported countries to **convene national dialogues on big data by inviting experts from the private sector or academia to outline methods and use cases.** This can help NSOs build and understand potential opportunities that align with priority business needs and garner initial buy-in with staff and senior executives. **Big data champions or sponsors** in the NSO and more broadly in government or the private sector that have influence and **can champion the cause and bring others on-board can also help raise awareness, secure buy-in, and build consensus.** Early discussions with technical partners with expertise in big data as well as other peer organizations who are more advanced on their big data journey can also raise awareness of potential applications. Advice suggests that it is important to **demonstrate real world case studies, and identifying one or two "low hanging fruit" with clear methods and applications and that use open access data sources can help to get things started** (e.g., price statistics or web scraping).

5.2 Building a Business Case and Aligning Incentives

Moving through the pre-adoption to early implementation stage requires the **elaboration of a sound business case for piloting big data applications for national statistics**, generally starting with one or two promising priority applications. A range of partners is often necessary in developing the business case, including NSOs, data providers, regulators, technical service providers, privacy officers, legal counsel etc. Aligning incentives among these various partners requires a compelling value proposition, and a clear articulation of the benefits and responsibilities for each partner. Incentives for public-private big data partnerships can include reciprocity, research insights, reputation and building public relations, regulatory compliance, and a new revenue stream (Klein and Verhulst, 2017). **It is important to clearly articulate the technical project requirements, costs and benefits, and to create a win-win situation to encourage all partners to continue to commit resources to the project.** A financing source is needed to cover project costs, which can be sourced through government allocations or international donors or partners.

In terms of using proprietary data sources for official statistics, experiences in several countries suggest demonstrating social value has played a greater role in motivating private sector partners to engage in big data projects to date. However, this is changing due to increasing awareness in the private sector of the value of their data. Over time, there will likely be an increasing need for commercial companies to balance corporate social responsibility with sustainable business models that can support projects over the long-term. Understanding the incentives and operational contexts of each stakeholder is critical in building a value proposition and securing alignment of incentives across partners (DIAL and Data-Pop Alliance, 2021).

5.3 Navigating the Governance and Regulatory Landscape

A key challenge highlighted by NSOs for successful proof of concept projects is related to the legal and regulatory landscape. However, this is primarily related to privately-owned big data sources, such as mobile phone data where formalized data sharing agreements were required. **Privacy and security risks may arise in big data projects with the transfer, storage, and sharing of data between stakeholders, and the potential for the unauthorized disclosure of personal data.** Partnership agreements such as MoUs or commercial contracts are often needed to be negotiated with data providers and can be used more broadly to provide the framework for other forms of partnerships, including with analytical or technology providers. To help navigate the regulatory landscape and privacy requirements, early engagement should be undertaken with data privacy agencies, regulators, and officers, as well as legal counsel units within the NSO or government. Based on interviews, building trust with these partners was critical during the pre-adoption and early implementation stages and can lead to the development of standardized data sharing agreements or MoUs for different big data partnership arrangements. Examples of data sharing agreement templates are collated through the Contracts for Data Collaboration (C4DC) (**Box 5-1**), which can improve understanding of the legal conditions that enable effective data collaboration.

Box 5-1

Contracts for Data Collaboration (C4DC)

Data sharing agreements (DSAs) are written documents laying out terms by which data is shared between parties. They can be essential to managing collaborative relationships and navigating the associated risks. Yet negotiating DSAs often takes many months, and finalizing the agreement can be the most burdensome step in an entire collaboration. Countries are still gaining experience with when and how to create DSAs. Especially with the growing concentration of private data, there is a concern about DSAs becoming imbalanced.

The C4DC initiative works to provide policy resources documenting how DSAs are formed. An **online C4DC library** includes DSAs used in a variety of SDG related applications, along with accompanying case studies and reports. A review (Dahmm, 2020) of the example DSAs in the library revealed that a number do not actually address seemingly significant issues, suggesting current data sharing practices are generally not considering the full range of potential concerns. Importantly, the responsibility for managing data extends beyond the planned length of a project, so DSAs ought to address what will be done with the data once the collaboration concludes. With only a third of the library's DSAs actively addressing this issue, however, the sample suggests approaches are not yet common or standardized. Likewise, although the library demonstrates a range of ways limitations can be set on data use by recipients, less than two-thirds of the DSAs clarify use limitations. Additionally, less than half address data ownership, and less than a third lay out procedures in the event of a data breach.

The C4DC library also presents terms that future collaborators might consider when determining how to address these and other issues in their own DSAs. Yet it is important to recognize there is no "universal" DSA, as the circumstances faced by one team of data collaborators are often unique, and parties should use their own legal counsel when finalizing an agreement.

5.4 Building the Technical and Human Capabilities

The early implementation stage provides an opportunity to begin building the necessary human capabilities and infrastructure for big data applications. Analytical and technology partners are critical at this stage to provide technical support services, training, and advice to help complement and grow in-house capacity. Establishing relationships with national geoscience or mapping agencies, space agencies, academic and research institutes, and domain experts are essential for developing national capabilities. Furthermore, given the specialized nature of datasets and skillsets, NSOs will need to partner with multidisciplinary teams comprising subject matter experts, data scientists, statisticians, computer scientists, among others to harness big data sources.

5.5 Sustaining Big Data Partnerships

Moving from early implementation and pilot projects to proficiency and mature big data statistical applications takes time, commitment, and ongoing investment. Scaling up from pilot activities to production and sustaining these activities over the long-term will likely require NSOs to make significant investments in human capital and skills development. Appropriately skilled data scientists will be needed to collect, process, and analyze big data sources, and to integrate new data collection and production methods into their work programs. This will take time and require ongoing capacity-building, training, methodological advancements, and financing. These efforts will also need to be backed by ongoing big data programs that are planned and budgeted with secure funding. Data providers, analytical partners, technology partners, and donors will therefore be critical in supporting NSOs to cross the chasm towards big data proficiency.

Governments will also need to invest considerable efforts in establishing an authorizing framework that sets out the rules for access to and use of big data that provides transparency with regard to the roles of the data custodians and the relevant government offices (UNSD and UNECE, 2015). Putting in place effective legislative frameworks that enable data sharing while safe-guarding privacy will take time. Ongoing engagement with legal counsel and privacy organizations is needed to help develop an appropriate and lasting legal framework.

Even for organizations that have reached big data proficiency or maturity, partnerships will still play an important role in data collection and production, particularly with providers of big data sources as well as technology partners providing bespoke infrastructure (e.g., cloud computing and geospatial platforms), analytical partners supporting methodological advancements and algorithms, and legal and regulatory partners providing counsel and enforcing data protection laws.

6. Conclusion

This report confirms the strong value proposition of big data in addressing the unprecedented monitoring challenge associated with the SDGs. It finds that national experience in using big data in statistical production has matured over the past decade, and cites a number of recent examples of countries experimenting with big data to support national SDGs monitoring. This is being enabled by rapid methodological advancements in the research community, in particular the use of open access satellite or EO data sources to generate new datasets can serve a range of monitoring objectives. Given the pressing need to fill data gaps to support SDGs monitoring, it will be important to harness all promising innovations across the data ecosystem. However, this must be done in a way that ensures national ownership and supports NSOs in their central coordinating role.

Recent examples of global and national applications of big data sources to support SDG monitoring highlight a complex ecosystem involving governments, private enterprises, academia, international organizations, civil society, and technology service providers. This new data ecosystem challenges the traditional data production model, requiring NSOs to tap new data sources, adopt new techniques, utilize new platforms and work with new partners. However, it is important to remember datasets derived from big data are a proxy measure or estimate providing a certain likelihood or confidence level that a phenomenon occurs. Innovative sources of data are therefore not perfect substitutes for traditional data sources on which many official statistics are based. It is important that estimation methods incorporating big data do not aim to replace conventional methods, and the integration of conventional and innovative sources can help to ensure accuracy and reliability while improving granularity, coverage, and timeliness.

Building awareness and developing maturity in skills and capabilities more widely in NSOs will be critical to mobilizing opportunities, given the broad utility of big data sources for SDG monitoring. These capabilities clearly exist in a range of partners globally, who will be critical in enabling NSOs to harness the potential of big data for SDGs monitoring.

NSOs are at very different stages in their big data journey and have different levels of maturity in terms of their big data awareness and literacy, human capabilities and infrastructure, and governance and regulatory frameworks. As such, there is no one-size-fits-all solution and partners can play many different roles. The big data and partnerships maturity model presented in the report provides an intuitive framework to help NSOs to understand their current state and progress in terms of deriving value from big data. During the journey from nascent maturity through to big data proficiency, NSOs will need to work with and benefit from a range of partners and partnership models. Emerging experiences from pioneer countries highlight a number of practical steps NSOs can take to develop their big data maturity and partners who can help along the journey.

References

ABATZOGLOU, J. T., DOBROWSKI, S. Z., PARKS, S. A. & HEGEWISCH, K. C. 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific data,* 5**,** 170191. https://doi.org/10.1038/sdata.2017.191

ADB 2020. Mapping poverty through data integration and artificial intelligence. *A Special Supplement of the Key Indicators for Asia and the Pacific 2020.* Manila: Asian Development Bank.

ALI, A., QADIR, J., UR RASOOL, R., SATHIASEELAN, A., ZWITTER, A. & CROWCROFT, J. 2016. Big data for development: applications and techniques. *Big Data Analytics,* 1**,** 2. https://doi.org/10.1186/s41044-016-0002-4

ALLEN, C., SMITH, M., RABIEE, M. & DAHMM, H. 2021. A review of scientific advancements in datasets derived from big data for monitoring the Sustainable Development Goals. *Sustainability Science*. https://doi.org/10.1007/s11625-021-00982-3

ANDERSSON, M., HALL, O. & ARCHILA, M. F. 2019. How Data-Poor Countries Remain Data Poor: Underestimation of Human Settlements in Burkina Faso as Observed from Nighttime Light Data. *ISPRS International Journal of Geo-Information,* 8**,** 498. https://doi.org/10.3390/ijgi8110498

ANDREANO, M. S., BENEDETTI, R., PIERSIMONI, F. & SAVIO, G. 2020. Mapping Poverty of Latin American and Caribbean Countries from Heaven Through Night-Light Satellite Images. *Social Indicators Research***,** 1-30. https://doi.org/10.1007/s11205-020-02267-1

ARDERNE, C., ZORN, C., NICOLAS, C. & KOKS, E. 2020. Predictive mapping of the global power system using open data. *Scientific data,* 7**,** 1-12. https://doi.org/10.1038/s41597-019-0347-4

AVITABILE, V., HEROLD, M., HEUVELINK, G. B., LEWIS, S. L., PHILLIPS, O. L., ASNER, G. P., ARMSTON, J., ASHTON, P. S., BANIN, L. & BAYOL, N. 2016. An integrated pan□tropical biomass map using multiple reference datasets. *Global change biology,* 22**,** 1406-1420. https://doi.org/10.1111/gcb.13139

AVTAR, R., AGGARWAL, R., KHARRAZI, A., KUMAR, P. & KURNIAWAN, T. A. 2020. Utilizing geospatial information to implement SDGs and monitor their Progress. *Environmental Monitoring and Assessment,* 192**,** 35. https://doi.org/10.1007/s10661-019-7996-9

BANTANG, J., SAN BUENAVENTURA, A. & GARRAEZ, J. 2020. Going Beyond Measuring the Rural Access Index in the Philippines. Manila: Philippine Statistics Authority.

BASTIAANSSEN, W. G. & STEDUTO, P. 2017. The water productivity score (WPS) at global and regional level: Methodology and first results from remote sensing measurements of wheat, rice and maize. *Science of the Total Environment,* 575**,** 595-611. http://dx.doi.org/10.1016/j.scitotenv.2016.09.032

BECKER-RESHEF, I., FRANCH, B., BARKER, B., MURPHY, E., SANTAMARIA-ARTIGAS, A., HUMBER, M., SKAKUN, S. & VERMOTE, E. 2018. Prior season crop type masks for winter wheat yield forecasting: a US case study. *Remote Sensing,* 10**,** 1659. https://doi.org/10.3390/rs10101659

BECKER-RESHEF, I., JUSTICE, C., BARKER, B., HUMBER, M., REMBOLD, F., BONIFACIO, R., ZAPPACOSTA, M., BUDDE, M., MAGADZIRE, T. & SHITOTE, C. 2020. Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. *Remote Sensing of Environment,* 237**,** 111553. https://doi.org/10.1016/j.rse.2019.111553

BELOCONI, A., CHRYSOULAKIS, N., LYAPUSTIN, A., UTZINGER, J. & VOUNATSOU, P. 2018. Bayesian geostatistical modelling of PM10 and PM2. 5 surface level concentrations in Europe using high-resolution satellite-derived products. *Environment international,* 121**,** 57-70. https://doi.org/10.1016/j.envint.2018.08.041

BIAN, J., LI, A., LEI, G., ZHANG, Z. & NAN, X. 2020. Global high-resolution mountain green cover index mapping based on Landsat images and Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing,* 162**,** 63-76. https://doi.org/10.1016/j.isprsjprs.2020.02.011

BIERMANN, L., CLEWLEY, D., MARTINEZ-VICENTE, V. & TOPOUZELIS, K. 2020. finding plastic patches in coastal Waters using optical Satellite Data. *Scientific reports,* 10**,** 1-10. https://doi.org/10.1038/s41598-020-62298-z

BLUMENSTOCK, J. 2018. Don't forget people in the use of big data for development. Nature Publishing Group.

BURKHARD, K., SCHLATTMANN, A. & NEUENDORF, F. 2018. APPLICATION OF REMOTE SENSING AND BIG DATA FOR GLOBAL WATER USE SUSTAINABILITY ASSESSMENT AND MONITORING.

BURSTEIN, R., HENRY, N. J., COLLISON, M. L. & ET AL 2019. Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. *Nature,* 574**,** 353-358. https://doi.org/10.1038/s41586-019-1545-0

CAMPBELL, J., NEUNER, J., SEE, L., FRITZ, S., FRAISL, D., ESPEY, J. & KIM, A. 2020. The role of combining national official statistics with global monitoring to close the data gaps in the environmental SDGs. *Statistical Journal of the IAOS***,** 1-11.

CAZAREZ-GRAGEDA, K. & ZOUGBEDE, K. 2019. National SDG Review: data challenges and opportunities. Paris: Paris21 and Partners for Review.

CHIARELLI, D. D., PASSERA, C., ROSA, L., DAVIS, K. F., D'ODORICO, P. & RULLI, M. C. 2020. The green and blue crop water requirement WATNEEDS model and its global gridded outputs. *Scientific data,* 7**,** 1-9. https://doi.org/10.1038/s41597-020-00612-0

CHINESE ACADEMY OF SCIENCE 2020. Big Earth Data in Support of the Sustainable Development Goals. Beijing: Chinese Academy of Science.

CORBANE, C., MARTINO, P., PANAGIOTIS, P., ANETA, F. J., MICHELE, M., SERGIO, F., MARCELLO, S., DANIELE, E., GUSTAVO, N. & THOMAS, K. 2020. The grey-green divide: multi-temporal analysis of greenness across 10,000 urban centres derived from the Global Human Settlement Layer (GHSL). *International Journal of Digital Earth,* 13**,** 101-118. https://doi.org/10.1080/17538947.2018.1530311

COWIE, A. L., ORR, B. J., SANCHEZ, V. M. C., CHASEK, P., CROSSMAN, N. D., ERLEWEIN, A., LOUWAGIE, G., MARON, M., METTERNICHT, G. I. & MINELLI, S. 2018. Land in balance: The scientific conceptual framework for Land Degradation Neutrality. *Environmental Science & Policy,* 79**,** 25-35. https://doi.org/10.1016/j.envsci.2017.10.011

CUARESMA, J. C., FENGLER, W., KHARAS, H., BEKHTIAR, K., BROTTRAGER, M. & HOFER, M. 2018. Will the Sustainable Development Goals be fulfilled? Assessing present and future global poverty. *Palgrave Communications,* 4**,** 1-8.

DAAS, P. J., DE BROE, S. & VAN MEETEREN, M. 2016. Center for Big Data Statistics at Statistics Netherlands. Netherlands: Statistics Netherlands.

DAAS, P. J., PUTS, M. J., BUELENS, B. & VAN DEN HURK, P. A. 2015. Big data as a source for official statistics. *Journal of Official Statistics,* 31**,** 249-262. https://doi.org/10.1515/jos-2015-0016

DAHMM, H. 2020. Laying the Foundation for Effective Partnerships: An Examination of Data Sharing Agreements. New York: Center for Open Science.

DANE 2018. Using Earth Observation data for calculating SDG indicators in Colombia. *Presentation to the 8th Meeting of the IAEG-SDGs.* Government of Colombia.

DIAL & DATA-POP ALLIANCE 2021. MD4D Handbook. *An interactive resource to gain knowledge and skills for designing and implementing mobile data for development (MD4D) projects.* Digital Impact Alliance.

DUNNETT, S., SORICHETTA, A., TAYLOR, G. & EIGENBROD, F. 2020. Harmonised global datasets of wind and solar farm locations and power. *Scientific Data,* 7**,** 1-12. https://doi.org/10.1038/s41597-020-0469-8

ESSC 2013. Scheveningen Memorandum on Big Data and Official Statistics. Scheveningen: European Statistical System Committee.

ESSC 2018. Bucharest Memorandum on Official Statistics in a Datafield Society (Trusted Smart Statistics). Bucharest: European Statistical System Committee.

FALCHETTA, G., PACHAURI, S., PARKINSON, S. & BYERS, E. 2019. A high-resolution gridded dataset to assess electrification in sub-Saharan Africa. *Scientific data,* 6**,** 1-9. https://doi.org/10.1038/s41597-019-0122-6

FATEHKIA, M., KASHYAP, R. & WEBER, I. 2018. Using Facebook ad data to track the global digital gender gap. *World Development,* 107**,** 189-209. https://doi.org/10.1016/j.worlddev.2018.03.007

FATEHKIA, M., TINGZON, I., ORDEN, A., SY, S., SEKARA, V., GARCIA-HERRANZ, M. & WEBER, I. 2020. Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science,* 9**,** 22. https://doi.org/10.1140/epjds/s13688-020-00235-w

FLORESCU, D., KARLBERG, M., REIS, F., DEL CASTILLO, P. R., SKALIOTIS, M. & WIRTHMANN, A. Will 'big data'transform official statistics. European Conference on the Quality of Official Statistics., 2014 Vienna, Austria. 2-5.

FORKUOR, G., ZOUNGRANA, J.-B. B., DIMOBE, K., OUATTARA, B., VADREVU, K. P. & TONDOH, J. E. 2020. Above-ground biomass mapping in West African dryland forest using Sentinel-1 and 2 datasets-A case study. *Remote Sensing of Environment,* 236**,** 111496. https://doi.org/10.1016/j.rse.2019.111496

FRAISL, D., CAMPBELL, J., SEE, L., WEHN, U., WARDLAW, J., GOLD, M., MOORTHY, I., ARIAS, R., PIERA, J. & OLIVER, J. L. 2020. Mapping citizen science contributions to the UN sustainable development goals. *Sustainability Science,* 15**,** 1735-1751. https://doi.org/10.1007/s11625-020-00833-7

FUNK, C., PETERSON, P., LANDSFELD, M., PEDREROS, D., VERDIN, J., SHUKLA, S., HUSAK, G., ROWLAND, J., HARRISON, L. & HOELL, A. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific data,* 2**,** 1-21. https://doi.org/10.1038/sdata.2015.66

GANDOMI, A. & HAIDER, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management,* 35**,** 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

GARCIA, D., KASSA, Y. M., CUEVAS, A., CEBRIAN, M., MORO, E., RAHWAN, I. & CUEVAS, R. 2018. Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences,* 115**,** 6958-6963. https://doi.org/10.1073/pnas.1717781115

GBD 2016 SDG COLLABORATORS 2017. Measuring progress and projecting attainment on the basis of past trends of the health-related Sustainable Development Goals in 188 countries: an analysis from the Global Burden of Disease Study 2016. *The Lancet,* 390**,** 1423-1459. https://doi.org/10.1016/S0140-6736(17)32336-X

GBD 2019 DISEASES AND INJURIES COLLABORATORS 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet,* 396**,** 1204-1222. https://doi.org/10.1016/S0140-6736(20)30925-9

GBD 2019 UNIVERSAL HEALTH COVERAGE COLLABORATORS 2020. Measuring universal health coverage based on an index of effective coverage of health services in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet,* 396**,** 1250-1284. https://doi.org/10.1016/S0140-6736(20)30750-9

GIULIANI, G., MAZZETTI, P., SANTORO, M., NATIVI, S., VAN BEMMELEN, J., COLANGELI, G. & LEHMANN, A. 2020. Knowledge generation using satellite earth observations to support sustainable development goals (SDG): A use case on Land degradation. *International Journal of Applied Earth Observation and Geoinformation,* 88**,** 102068. https://doi.org/10.1016/j.jag.2020.102068

GOSLING, J., JONES, M. I., ARNELL, A., WATSON, J. E., VENTER, O., BAQUERO, A. C. & BURGESS, N. D. 2020. A global mapping template for natural and modified habitat across terrestrial Earth. *Biological Conservation,* 250**,** 108674. https://doi.org/10.1016/j.biocon.2020.108674

GRAETZ, N., FRIEDMAN, J., OSGOOD-ZIMMERMAN, A., BURSTEIN, R., BIEHL, M. H., SHIELDS, C., MOSSER, J. F., CASEY, D. C., DESHPANDE, A. & EARL, L. 2018. Mapping local variation in educational attainment across Africa. *Nature,* 555**,** 48-53. https://doi.org/10.1038/nature25761

HAKIMDAVAR, R., HUBBARD, A., POLICELLI, F., PICKENS, A., HANSEN, M., FATOYINBO, T., LAGOMASINO, D., PAHLEVAN, N., UNNINAYAR, S. & KAVVADA, A. 2020. Monitoring water-related ecosystems with Earth observation data in support of Sustainable Development Goal (SDG) 6 reporting. *Remote Sensing,* 12**,** 1634. https://doi.org/10.3390/rs12101634

HALPER, F. & KRISHNAN, K. 2013. TDWI Big Data Maturity Model Guide. The Data Warehousing Institute (TDWI).

HAMMER, C., KOSTROCH, M. D. C. & QUIROS, M. G. 2017. *Big Data: Potential, Challenges and Statistical Implications*, International Monetary Fund.

HANSEN, A., BARNETT, K., JANTZ, P., PHILLIPS, L., GOETZ, S. J., HANSEN, M., VENTER, O., WATSON, J. E., BURNS, P. & ATKINSON, S. 2019. Global humid tropics forest structural condition and forest structural integrity maps. *Scientific Data,* 6**,** 1-12. https://doi.org/10.1038/s41597-019-0214-3

HANSEN, M. C., KRYLOV, A., TYUKAVINA, A., POTAPOV, P. V., TURUBANOVA, S., ZUTTA, B., IFO, S., MARGONO, B., STOLLE, F. & MOORE, R. 2016. Humid tropical forest disturbance alerts using Landsat data. *Environmental Research Letters,* 11**,** 034008. doi:10.1088/1748-9326/11/3/034008

HEEGE, T., SCHENK, K. & WILHELM, M.-L. 2019. Water Quality Information for Africa from Global Satellite Based Measurements: The Concept Behind the UNESCO World Water Quality Portal. *Embedding Space in African Society.* Springer.

HENGL, T., MENDES DE JESUS, J., HEUVELINK, G. B., RUIPEREZ GONZALEZ, M., KILIBARDA, M., BLAGOTIĆ, A., SHANGGUAN, W., WRIGHT, M. N., GENG, X. & BAUER-MARSCHALLINGER, B. 2017. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one,* 12**,** e0169748. https://doi.org/10.1371/journal.pone.0169748

HERSH, J., ENGSTROM, R. & MANN, M. 2020. Open data for algorithms: mapping poverty in Belize using open satellite derived features and machine learning. *Information Technology for Development***,** 1-30. https://doi.org/10.1080/02681102.2020.1811945

IAEG-SDGS 2020. Tier Classification for Global SDG Indicators. New York: Inter-Agency and Expert Group on SDG Indicators.

IVAN, K., HOLOBÂCĂ, I.-H., BENEDEK, J. & TÖRÖK, I. 2020. Potential of Night-Time Lights to Measure Regional Inequality. *Remote Sensing,* 12**,** 33. https://doi.org/10.3390/rs12010033

JEAN, N., BURKE, M., XIE, M., DAVIS, W. M., LOBELL, D. B. & ERMON, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science,* 353**,** 790-794. http://doi.org/10.1126/science.aaf7894

JENSEN, T., SEERUP HASS, F., SEAM AKBAR, M., HOLM PETERSEN, P. & JOKAR ARSANJANI, J. 2020. Employing Machine Learning for Detection of Invasive Species using Sentinel-2 and AVIRIS Data: The Case of Kudzu in the United States. *Sustainability,* 12**,** 3544. https://doi.org/10.3390/su12093544

JONES, E. R., VAN VLIET, M. T., QADIR, M. & BIERKENS, M. F. 2020. Spatially-explicit estimates of global wastewater production, collection, treatment and re-use. *Earth System Science Data Discussions***,** 1-29. https://doi.org/10.5194/essd-2020-156

JUNG, M., DAHAL, P. R., BUTCHART, S. H., DONALD, P. F., DE LAMO, X., LESIV, M., KAPOS, V., RONDININI, C. & VISCONTI, P. 2020. A global map of terrestrial habitat types. *Scientific data,* 7**,** 1-8. https://doi.org/10.1038/s41597-020-00599-8

JUVENTIA, S. D., JONES, S. K., LAPORTE, M.-A., REMANS, R., VILLANI, C. & ESTRADA-CARMONA, N. 2020. Text mining national commitments towards agrobiodiversity conservation and use. *Sustainability,* 12**,** 715. https://doi.org/10.3390/su12020715

KASHYAP, R., FATEHKIA, M., AL TAMIME, R. & WEBER, I. 2020. Monitoring global digital gender inequality using the online populations of Facebook and Google. *Demographic Research,* 43**,** 779-816. http://doi.org/10.4054/DemRes.2020.43.27

KAVVADA, A., METTERNICHT, G., KERBLAT, F., MUDAU, N., HALDORSON, M., LALDAPARSAD, S., FRIEDL, L., HELD, A. & CHUVIECO, E. 2020. Towards delivering on the sustainable development goals using earth observations. Elsevier.

KICKBUSCH, I. & HANEFELD, J. 2017. Role for academic institutions and think tanks in speeding progress on sustainable development goals. *Bmj,* 358**,** j3519. https://doi.org/10.1136/bmj.j3519

KITCHIN, R. 2015. The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS,* 31**,** 471-481. http://dx.doi.org/10.2139/ssrn.2595075

KLEIN, T. & VERHULST, S. 2017. Access to New Data Sources for Statistics: Business Models and Incentives for the Corporate Sector. Paris: Paris21.

LANEY, D. 2001. 3D data management: Controlling data volume, velocity and variety. *META group research note,* 6**,** 1.

LAWAL, O. & ANYIAM, F. E. 2019. Modelling geographic accessibility to primary health care facilities: Combining open data and geospatial analysis. *Geo-Spatial Information Science,* 22**,** 174-184. https://doi.org/10.1080/10095020.2019.1645508

LE QUÉRÉ, C., JACKSON, R. B., JONES, M. W., SMITH, A. J., ABERNETHY, S., ANDREW, R. M., DE-GOL, A. J., WILLIS, D. R., SHAN, Y. & CANADELL, J. G. 2020. Temporary reduction in daily global CO 2 emissions during the COVID-19 forced confinement. *Nature Climate Change***,** 1-7. https://doi.org/10.1038/s41558-020-0797-x

LEVIN, N., ALI, S. & CRANDALL, D. 2018. Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study. *Applied geography,* 94**,** 1-17. https://doi.org/10.1016/j.apgeog.2018.03.001

LI, J., CHEN, W.-H., XU, Q., SHAH, N. & MACKEY, T. Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology. 2019 IEEE Global Humanitarian Technology Conference (GHTC), 2019. IEEE, 1-4.

LI, W., EL-ASKARY, H., LAKSHMI, V., PIECHOTA, T. & STRUPPA, D. 2020a. Earth Observation and Cloud Computing in Support of Two Sustainable Development Goals for the River Nile Watershed Countries. *Remote Sensing,* 12**,** 1391. https://doi.org/10.3390/rs12091391

LI, Y., YU, M., XU, M., YANG, J., SHA, D., LIU, Q. & YANG, C. 2020b. Big Data and Cloud Computing. *Manual of Digital Earth***,** 325. https://doi.org/10.1007/978-981-32-9915-3_9

LOCAL BURDEN OF DISEASE CHILD GROWTH FAILURE COLLABORATORS 2020. Mapping child growth failure across low-and middle-income countries. *Nature,* 577**,** 231. https://doi.org/10.1038/s41586-019-1878-8

LUIS, O. D. A. J., CARLOS, C. P. J. & ALFREDO, S. M. H. Open Data Cube for Natural Resources Mapping in Mexico. Proceedings of the 1st International Conference on Geospatial Information Sciences, 2019. 70-78.

MACFEELY, S. 2019. The big (data) bang: Opportunities and challenges for compiling SDG indicators. *Global Policy,* 10**,** 121-133. https://doi.org/10.1111/1758-5899.12595

MARCONCINI, M., METZ-MARCONCINI, A., ÜREYEN, S., PALACIOS-LOPEZ, D., HANKE, W., BACHOFER, F., ZEIDLER, J., ESCH, T., GORELICK, N. & KAKARLA, A. 2020. Outlining where humans live, the World Settlement Footprint 2015. *Scientific Data,* 7**,** 1-14. https://doi.org/10.1038/s41597-020-00580-5

MARCOVECCHIO, I., THINYANE, M., ESTEVEZ, E. & FILLOTTRANI, P. 2018. Capability Maturity Models as a Means to Standardize Sustainable Development Goals Indicators Data Production. *Journal of ICT Standardization,* 6**,** 217-244.

MAUSER, W., HANK, T., JAKSZTAT, T. & PROBST, E. 2018. VIRTUAL WATER VALUES-A PROJECT FOR GLOBAL AND REGIONAL ASSESSMENT OF AGRICULTURAL YIELDS AND WATER USE EFFICIENCY. *SCIENTIFIC PAPERS-SERIES E-LAND RECLAMATION EARTH OBSERVATION & SURVEYING ENVIRONMENTAL ENGINEERING,* 7**,** 192-197.

MCOWEN, C. J., WEATHERDON, L. V., VAN BOCHOVE, J.-W., SULLIVAN, E., BLYTH, S., ZOCKLER, C., STANWELL-SMITH, D., KINGSTON, N., MARTIN, C. S. & SPALDING, M. 2017. A global map of saltmarshes. *Biodiversity data journal*. https://doi.org/10.3897/BDJ.5.e11764

MENTIS, D., HOWELLS, M., ROGNER, H., KORKOVELOS, A., ARDERNE, C., ZEPEDA, E., SIYAL, S., TALIOTIS, C., BAZILIAN, M. & DE ROO, A. 2017. Lighting the World: the first application of an open source, spatial electrification tool (OnSSET) on Sub-Saharan Africa. *Environmental Research Letters,* 12**,** 085003. https://doi.org/10.1088/1748-9326/aa7b29

METTERNICHT, G., MUELLER, N. & LUCAS, R. 2020. Digital Earth for Sustainable Development Goals. *Manual of Digital Earth.* Singapore: Springer.

MEYER, M. F., LABOU, S. G., CRAMER, A. N., BROUSIL, M. R. & LUFF, B. T. 2020. The global lake area, climate, and population dataset. *Scientific Data,* 7**,** 1-12. https://doi.org/10.1038/s41597-020-0517-4

MITRI, G., NASRALLAH, G., GEBRAEL, K., NASSAR, M. B., ABOU DAGHER, M., NADER, M., MASRI, N. & CHOUEITER, D. 2019. Assessing land degradation and identifying potential sustainable land management practices at the subnational level in Lebanon. *Environmental monitoring and assessment,* 191**,** 567. https://doi.org/10.1007/s10661-019-7739-y

MONDAL, P., LIU, X., FATOYINBO, T. E. & LAGOMASINO, D. 2019. Evaluating Combinations of Sentinel-2 Data and Machine-Learning Algorithms for Mangrove Mapping in West Africa. *Remote Sensing,* 11**,** 2928. https://doi.org/10.3390/rs11242928

MONDAL, P., MCDERMID, S. S. & QADIR, A. 2020. A reporting framework for Sustainable Development Goal 15: Multi-scale monitoring of forest degradation using MODIS, Landsat and Sentinel data. *Remote Sensing of Environment,* 237**,** 111592. https://doi.org/10.1016/j.rse.2019.111592

MUELLER, N., LEWIS, A., ROBERTS, D., RING, S., MELROSE, R., SIXSMITH, J., LYMBURNER, L., MCINTYRE, A., TAN, P. & CURNOW, S. 2016. Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia. *Remote Sensing of Environment,* 174**,** 341-352. https://doi.org/10.1016/j.rse.2015.11.003

MURRAY, N. J., PHINN, S. R., DEWITT, M., FERRARI, R., JOHNSTON, R., LYONS, M. B., CLINTON, N., THAU, D. & FULLER, R. A. 2019. The global distribution and trajectory of tidal flats. *Nature,* 565**,** 222-225. https://doi.org/10.1038/s41586-018-0805-8

NHAMO, L., MAGIDI, J. & DICKENS, C. 2017. Determining wetland spatial extent and seasonal variations of the inundated area using multispectral remote sensing. *Water Sa,* 43**,** 543-552.

OECD 2018. Which Strategies for NSOs in the Digital Era? Towards 'Smart Data' Strategies. Paris: Organisation for Economic Cooperation and Development.

OPEN DATA WATCH 2018. The Data Value Chain: Moving from Production to Impact.

ORR, B., COWIE, A., CASTILLO SANCHEZ, V., CHASEK, P., CROSSMAN, N., ERLEWEIN, A., LOUWAGIE, G., MARON, M., METTERNICHT, G. & MINELLI, S. Scientific conceptual framework for land degradation neutrality. Bonn, Germany: United Nations Convention to Combat Desertification (UNCCD), 2017. 1-98.

OSGOOD-ZIMMERMAN, A., MILLEAR, A. I., STUBBS, R. W., SHIELDS, C., PICKERING, B. V., EARL, L., GRAETZ, N., KINYOKI, D. K., RAY, S. E. & BHATT, S. 2018. Mapping child growth failure in Africa between 2000 and 2015. *Nature,* 555**,** 41-47. https://doi.org/10.1038/nature25760

OSHRI, B., HU, A., ADELSON, P., CHEN, X., DUPAS, P., WEINSTEIN, J., BURKE, M., LOBELL, D. & ERMON, S. Infrastructure quality assessment in Africa using satellite imagery and deep learning. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018. 616-625.

OUKO, E., OMONDI, S., MUGO, R., WAHOME, A., KASERA, K., NKURUNZIZA, E., KIEMA, J., FLORES, A., ADAMS, E. C. & KURARU, S. 2020. Modeling Invasive Plant Species in Kenya's Northern Rangelands. *Frontiers in Environmental Science,* 8**,** 69. https://doi.org/10.3389/fenvs.2020.00069

PARIS21 2020. Use of Citizen-Generated Data for SDG Reporting in the Philippines: a case study. Paris: Paris21.

PARTNERSHIPS TASK TEAM 2014. Guidelines for the establishment and use of partnerships in big data projects for official statistics. *In:* SERVICES, T. T. O. T. H.-L. G. F. T. M. O. S. P. A. (ed.). Geneva: United Nations Economic Commission for Europe.

PEKEL, J.-F., COTTAM, A., GORELICK, N. & BELWARD, A. S. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature,* 540**,** 418-422. https://doi.org/10.1038/nature20584

PENG, J., DADSON, S., HIRPA, F., DYER, E., LEES, T., GONZALEZ MIRALLES, D., VICENTE-SERRANO, S. M. & FUNK, C. 2020. A pan-African high-resolution drought index dataset. *Earth System Science Data,* 12**,** 753-769. http://doi.org/10.5194/essd-12-753-2020

PICKENS, A. H., HANSEN, M. C., HANCHER, M., STEHMAN, S. V., TYUKAVINA, A., POTAPOV, P., MARROQUIN, B. & SHERANI, Z. 2020. Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time series. *Remote Sensing of Environment,* 243**,** 111792. https://doi.org/10.1016/j.rse.2020.111792

POKHRIYAL, N. & JACQUES, D. C. 2017. Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences,* 114**,** E9783-E9792. https://doi.org/10.1073/pnas.1700319114

RADERMACHER, W. J. 2018. Official statistics in the era of big data opportunities and threats. *International Journal of Data Science and Analytics,* 6**,** 225-231. https://doi.org/10.1007/s41060-018-0124-z

REIMSBACH-KOUNATZE, C. 2015. The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies. Paris: Organisation for Economic Cooperation and Development.

ROBIN, N., KLEIN, T. & JUTTING, J. 2016. Public-Private Partnerships for Statistics: Lessons Learned, Future Steps. Paris: Organisation for Economic Cooperation and Development.

SACHS, J., SCHMIDT-TRAUB, G., KROLL, C., LAFORTUNE, G., FULLER, G. & WOELM, F. 2020. The Sustainable Development Goals and COVID-19: Sustainable Development Report 2020. United Kingdom: Cambridge University Press.

SAURA, S., BERTZKY, B., BASTIN, L., BATTISTELLA, L., MANDRICI, A. & DUBOIS, G. 2019. Global trends in protected area connectivity from 2010 to 2018. *Biological conservation,* 238**,** 108183. https://doi.org/10.1016/j.biocon.2019.07.028

SAYRE, R., KARAGULLE, D., FRYE, C., BOUCHER, T., WOLFF, N. H., BREYER, S., WRIGHT, D., MARTIN, M., BUTLER, K. & VAN GRAAFEILAND, K. 2020. An assessment of the representation of ecosystems in global protected areas using new maps of World Climate Regions and World Ecosystems. *Global Ecology and Conservation,* 21**,** e00860. https://doi.org/10.1016/j.gecco.2019.e00860

SCANNAPIECO, M., VIRGILLITO, A. & ZARDETTO, D. 2013. Placing big data in official statistics: a big challenge. *New Techniques and Technologies in Statistics.* Italian National Institute of Statistics.

SCHAPHOFF, S., VON BLOH, W., RAMMIG, A., THONICKE, K., BIEMANS, H., FORKEL, M., GERTEN, D., HEINKE, J., JÄGERMEYR, J. & KNAUER, J. 2018. LPJmL4–a dynamic global vegetation model with managed land–Part 1: Model description. *Geoscientific Model Development,* 11**,** 1343-1375. http://doi.org/10.5194/gmd-11-1343-2018.

SCHIAVINA, M., MELCHIORRI, M., CORBANE, C., FLORCZYK, A. J., FREIRE, S., PESARESI, M. & KEMPER, T. 2019. Multi-Scale Estimation of Land-Use Efficiency (SDG 11.3. 1) across 25 Years Using Global Open and Free Data. *Sustainability,* 11**,** 5674. https://doi.org/10.3390/su11205674

SEE, L., FRITZ, S., MOORTHY, I., DANYLO, O., VAN DIJK, M. & RYAN, B. 2018. Using Remote Sensing and Geospatial Information for Sustainable Development. *In:* DESAI, R., KATO, H., KHARAS, H. & MCARTHUR, J. (eds.) *From Summits to Solutions: Innovations in Implementing the Sustainable Development Goals.* Washington, DC: Brookings Institution Press.

SHADDICK, G., THOMAS, M., MUDU, P., RUGGERI, G. & GUMY, S. 2020. Half the world's population are exposed to increasing air pollution. *NPJ Climate and Atmospheric Science,* 3**,** 1-5. https://doi.org/10.1038/s41612-020-0124-2

SHADDICK, G., THOMAS, M. L., AMINI, H., BRODAY, D., COHEN, A., FROSTAD, J., GREEN, A., GUMY, S., LIU, Y. & MARTIN, R. V. 2018. Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment. *Environmental science & technology,* 52**,** 9069-9078. https://doi.org/10.1021/acs.est.8b02864

SHEEHAN, E., MENG, C., TAN, M., UZKENT, B., JEAN, N., BURKE, M., LOBELL, D. & ERMON, S. Predicting economic development using geolocated Wikipedia articles. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019. 2698-2706.

SHIFFMAN, J. & SHAWAR, Y. R. 2020. Strengthening accountability of the global health metrics enterprise. *The Lancet,* 395**,** 1452-1456. https://doi.org/10.1016/S0140-6736(20)30416-5

SIVARAJAH, U., KAMAL, M. M., IRANI, Z. & WEERAKKODY, V. 2017. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research,* 70**,** 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

SMITS, J. & PERMANYER, I. 2019. The subnational human development database. *Scientific data,* 6**,** 190038. https://doi.org/10.1038/sdata.2019.38

SONG, X.-P., HANSEN, M. C., STEHMAN, S. V., POTAPOV, P. V., TYUKAVINA, A., VERMOTE, E. F. & TOWNSHEND, J. R. 2018. Global land change from 1982 to 2016. *Nature,* 560**,** 639-643. https://doi.org/10.1038/s41586-018-0411-9

STARK, T., WURM, M., ZHU, X. X. & TAUBENBÖCK, H. 2020. Satellite-Based Mapping of Urban Poverty with Transfer-Learned Slum Morphologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 13**,** 5251-5263. http://doi.org/10.1109/JSTARS.2020.3018862

STEELE, J. E., SUNDSØY, P. R., PEZZULO, C., ALEGANA, V. A., BIRD, T. J., BLUMENSTOCK, J., BJELLAND, J., ENGØ-MONSEN, K., DE MONTJOYE, Y.-A. & IQBAL, A. M. 2017. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface,* 14**,** 20160690. https://doi.org/10.1098/rsif.2016.0690

STOKES, E. C. & SETO, K. C. 2019. Characterizing urban infrastructural transitions for the Sustainable Development Goals using multi-temporal land, population, and nighttime light data. *Remote Sensing of Environment,* 234**,** 111430. https://doi.org/10.1016/j.rse.2019.111430

STRUIJS, P., BRAAKSMA, B. & DAAS, P. J. 2014. Official statistics and big data. *Big Data & Society,* 1. https://doi.org/10.1177/2053951714538417

TAM, S.-M. & VAN HALDEREN, G. 2020. The five V's, seven virtues and ten rules of big data engagement for official statistics. *Statistical Journal of the IAOS,* 36**,** 423-433. http://doi.org/10.3233/SJI-190595

TAM, S. M. & CLARKE, F. 2015. Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review,* 83**,** 436-448. https://doi.org/10.1111/insr.12105

TUSTING, L. S., BISANZIO, D., ALABASTER, G., CAMERON, E., CIBULSKIS, R., DAVIES, M., FLAXMAN, S., GIBSON, H. S., KNUDSEN, J. & MBOGO, C. 2019. Mapping changes in housing in sub-Saharan Africa from 2000 to 2015. *Nature,* 568**,** 391-394. https://doi.org/10.1038/s41586-019-1050-5

UN STATISTICS DIVISION 2017. Guidelines and Best Practices on Data Flows and Global Data Reporting for Sustainable Development Goals. New York: Statistics Division.

UNESCAP 2020. Incorporating Non-Traditional Data Sources into Official Statistics. Bangkok: United Nations Economic and Social Commission for Asia and the Pacific.

UNESCAP 2021. Big Data for the SDGs. *Country examples in compiling SDG indicators using non-traditional sources.* Bangkok: United Nations Economic and Social Commission for Asia and the Pacific.

UNGA 2015. Transforming our world: the 2030 Agenda for Sustainable Development, outcome document of the United Nations summit for the adoption of the post-2015 agenda. *RES/A/70/L.1.* New York: United Nations General Assembly.

UNGWG FOR BIG DATA 2019. Handbook on the Use of Mobile Phone Data for Official Statistics. New York: UN Global Working Group on Big Data for Official Statistics.

UNITED NATIONS GENERAL ASSEMBLY 2017. Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development. New York.

UNITED NATIONS STATISTICAL COMMISSION 2014. Big data and modernization of statistical systems;. *Report of the Secretary-General. E/CN. 3.2014/11 of the forty-fifth session of UNSC 4-7 March 2014.* New York: United Nations.

UNSD & UNECE 2015. Results of the UNSD/UNECE Survey on organizational context and individual projects of Big Data. New York: United Nations Statistics Division.

UNSD & WORLD BANK 2020. Monitoring the State of Statistical Operations under the COVID-19 Pandemic. *Highlights from the Third Round of a Global COVID-19 Survey of National Statistical Offices.* New York: United Nations.

VAN DEVENTER, H., VAN NIEKERK, L., ADAMS, J., DINALA, M. K., GANGAT, R., LAMBERTH, S. J., LÖTTER, M., MBONA, N., MACKAY, F. & NEL, J. L. 2020. National Wetland Map 5: An improved spatial extent and representation of inland aquatic and estuarine ecosystems in South Africa. *Water SA,* 46**,** 66-79.

WANG, J., WEI, H., CHENG, K., OCHIR, A., DAVAASUREN, D., LI, P., CHAN, F. K. S. & NASANBAT, E. 2020. Spatio-temporal pattern of land degradation from 1990 to 2015 in Mongolia. *Environmental Development***,** 100497. https://doi.org/10.1016/j.envdev.2020.100497

WANG, X., SUTTON, P. C. & QI, B. 2019. Global Mapping of GDP at 1 km2 Using VIIRS Nighttime Satellite Imagery. *ISPRS International Journal of Geo-Information,* 8**,** 580. https://doi.org/10.3390/ijgi8120580

WATMOUGH, G. R., MARCINKO, C. L., SULLIVAN, C., TSCHIRHART, K., MUTUO, P. K., PALM, C. A. & SVENNING, J.-C. 2019. Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences,* 116**,** 1213-1218. https://doi.org/10.1073/pnas.1812969116

WEBER, I., KASHYAP, R. & ZAGHENI, E. 2018. Using advertising audience estimates to improve global development statistics. *Itu Journal: Ict Discoveries,* 1.

WEISE, K., HÖFER, R., FRANKE, J., GUELMAMI, A., SIMONSON, W., MURO, J., O'CONNOR, B., STRAUCH, A., FLINK, S. & EBERLE, J. 2020. Wetland extent tools for SDG 6.6. 1 reporting from the Satellite-based Wetland Observation Service (SWOS). *Remote Sensing of Environment,* 247**,** 111892. https://doi.org/10.1016/j.rse.2020.111892

WEISS, D., NELSON, A., VARGAS-RUIZ, C., GLIGORIĆ, K., BAVADEKAR, S., GABRILOVICH, E., BERTOZZI-VILLA, A., ROZIER, J., GIBSON, H. & SHEKEL, T. 2020. Global maps of travel time to healthcare facilities. *Nature Medicine***,** 1-4. https://doi.org/10.1038/s41591-020-1059-1

WEISS, D. J., NELSON, A., GIBSON, H., TEMPERLEY, W., PEEDELL, S., LIEBER, A., HANCHER, M., POYART, E., BELCHIOR, S. & FULLMAN, N. 2018. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature,* 553**,** 333-336. https://doi.org/10.1038/nature25181

WIGLEY, A., TEJEDOR-GARAVITO, N., ALEGANA, V., CARIOLI, A., RUKTANONCHAI, C. W., PEZZULO, C., MATTHEWS, Z., TATEM, A. & NILSEN, K. 2020. Measuring the availability and geographical accessibility of maternal health services across sub-Saharan Africa. *BMC medicine,* 18**,** 1-10. https://doi.org/10.1186/s12916-020-01707-6

WORLD BANK 20201. Data for Better Lives. *World Development Report 2021.* Washington: World Bank.

WURM, M., STARK, T., ZHU, X. X., WEIGAND, M. & TAUBENBÖCK, H. 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing,* 150**,** 59-69. https://doi.org/10.1016/j.isprsjprs.2019.02.006

YEH, C., PEREZ, A., DRISCOLL, A., AZZARI, G., TANG, Z., LOBELL, D., ERMON, S. & BURKE, M. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications,* 11**,** 1-11. https://doi.org/10.1038/s41467-020-16185-w

ZAUSSINGER, F., DORIGO, W., GRUBER, A., TARPANELLI, A., FILIPPUCCI, P. & BROCCA, L. 2019. Estimating irrigation water use over the contiguous United States by combining satellite and reanalysis soil moisture data. *Hydrology and earth system sciences,* 23**,** 897-923.

ZHOU, Y., LI, X., ASRAR, G. R., SMITH, S. J. & IMHOFF, M. 2018. A global record of annual urban dynamics (1992–2013) from nighttime lights. *Remote Sensing of Environment,* 219**,** 206-220. https://doi.org/10.1016/j.rse.2018.10.015

Appendix 1.

Use Cases for Big Data Sources to Support SDGs Monitoring – By Goal

Concerning **SDG1** on poverty, the 12 publications reviewed demonstrate new big data methods for mapping poverty rates (indicators 1.1.1, 1.2.1), multidimensional poverty (1.2.2), and access to basic services (1.4.1). For poverty indicators, the general objective of the studies is to demonstrate new methods using big data (e.g., satellite imagery, night-time lights (NTL), and OpenStreetMap (OSM)) to predict sub-national estimates of household wealth (e.g., Demographic Health Survey (DHS) Wealth Index) at various scales, including regionally for Latin America (Andreano et al., 2020) and Africa (Yeh et al., 2020), and for selected countries in Africa and Asia (Watmough et al., 2019, Hersh et al., 2020, Steele et al., 2017, Jean et al., 2016). The methods used are also now being applied in pilot case studies on small area estimate poverty mapping in several countries (**Box 3-2**). Some new studies also include other novel sources of big data, including Facebook marketing and user data (Fatehkia et al., 2020), Wikipedia articles (Sheehan et al., 2019), and call detail records (CDRs) (Steele et al., 2017). All of these studies use different machine learning (ML) regression and regularization methods (e.g., classification trees, random forest, object-based analysis, fractional regression, gaussian process regression, ensemble models, etc.), as well as deep learning (DL) methods (e.g., convolutional neural networks (CNNs)). An alternative approach is demonstrated by Cuaresma et al. (2018) who report results from a global study using a macro-econometric model and administrative and survey data to provide real-time predictions of poverty, which are made publicly available via the World Poverty Clock. Several other studies also make their datasets and code available through GitHub (Hersh et al., 2020, Yeh et al., 2020, Jean et al., 2016) or other repositories (Fatehkia et al., 2020). Collaborators include UN organizations, the World Bank, Brookings Institution, the International Institute for Applied Systems Analysis (IIASA), and Flowminder Foundation. Overall, around half of these datasets are provided open access.

Additionally, Pokhriyal and Jacques (2017) propose a computational framework to accurately predict the Global Multidimensional Poverty Index (MPI) (related to 1.2.2) in Senegal using environmental data and CDRs. In relation to 1.4.1, Weiss et al. (2018) develop and validate a global map that quantifies travel time to cities (2015 baseline) at a high spatial resolution, using a geospatial modelling framework. Collaborators include Orange Telecom, the Bill and Melinda Gates Foundation (BMGF), the European Commission Joint Research Centre (EC-JRC), Google, and the Malaria Atlas Project.

For **SDG 2** on hunger, studies cover aspects relating to health and nutrition (indicators 2.2.1, 2.2.2), as well as agriculture and food security (indicators 2.1.2, 2.3.1, 2.4.1). Recent studies by the Global Burden of Disease (GBD) collaborators (**Box 3-3**) apply Bayesian geostatistical modelling using a range of survey data, as well as big data covariates to report on child growth failure (Local Burden of Disease Child Growth Failure Collaborators, 2020, GBD 2016 SDG Collaborators, 2017, Osgood-Zimmerman et al., 2018) with all datasets and code provided open access. Other studies use bespoke geospatial models combined with satellite and climate or meteorological datasets to report on crop conditions (Becker-Reshef et al., 2020), agricultural yields (Becker-Reshef et al., 2018, Mauser et al., 2018), drought conditions, and crop water requirements (Peng et al., 2020, Chiarelli et al., 2020, Schaphoff et al., 2018), the latter of which provide datasets open access. Collaborators include EC-JRC, the World Food Program (WFP), the Food and Agricultural Organization of the UN (FAO), the United States Geological Survey (USGS), the Group on Earth Observations (GEO), the National Aeronautics and Space Administration (NASA), and the UK Space Agency (UKSA).

Compared to other goals, **SDG 3** on health had the largest coverage (75%) of official SDG indicators. As with the nutrition indicators above, this has largely been delivered through international collaboration under the GBD (**Box 3-3**) which contributes data on the incidence, prevalence, and mortality for a mutually exclusive and collectively exhaustive list of diseases and injuries (GBD 2019 Diseases and Injuries Collaborators, 2020). The annual GBD studies use geostatistical modelling that make inferences from spatially correlated phenomena to produce high spatial resolution estimates of SDG 3 indicators as well as health- or mortality-related indicators from several other SDGs (GBD 2019 Diseases and Injuries Collaborators, 2020, GBD 2016 SDG Collaborators, 2017, Burstein et al., 2019). Recently, several new studies have focused on the monitoring of access to healthcare, including through the GBD (GBD 2019 Universal Health Coverage Collaborators, 2020), as well as other studies at various scales (Weiss et al., 2020, Lawal and Anyiam, 2019, Wigley et al., 2020). All datasets for GBD studies and other global-scales studies are made open access and are funded through the BMGF and include collaboration with the World Health Organization (WHO).

Studies of relevance for **SDG 4** on education include geostatistical modelling using survey data and spatial covariates to produce an open access dataset on variation in educational attainment (i.e. average years of schooling; gender parity) across Africa (Graetz et al., 2018), as well as the development of the open access global Sub-national Human Development Index database, which incorporates average years of schooling (Smits and Permanyer, 2019). Collaborators include the BMGF and the European Research Council.

Several global studies use social media and tracking data from Facebook, Google, Twitter, and LinkedIn to map digital gender gaps of relevance for **SDG 5** (5.b.1) (Garcia et al., 2018, Kashyap et al., 2020, Weber et al., 2018, Fatehkia et al., 2018). The only dataset made open access is by Garcia et al. (2018). Collaborators include the Data2X initiative of the UN Foundation.

In addition to indicators on water access and sanitation (6.1.1, 6.2.1) reported through the GBD studies (GBD 2016 SDG Collaborators, 2017), other publications of relevance for **SDG 6** on water and sanitation include global spatially explicit estimates on wastewater production, treatment, and reuse (6.3.1) (Jones et al., 2020), global water quality monitoring using satellite data (6.3.2) (Heege et al., 2019), several studies using geospatial modelling on water use efficiency and crop productivity (6.4.1) (Zaussinger et al., 2019, Burkhard et al., 2018, Bastiaanssen and Steduto, 2017), a study applying deep learning to measure water stress in the Nile Delta region (6.4.2) (Li et al., 2020a), as well as several studies using satellite data, geospatial modelling, and machine learning methods to derive surface water extent and change dynamics (6.6.1) at global (Pekel et al., 2016, Pickens et al., 2020, Meyer et al., 2020) or other scales (Hakimdavar et al., 2020, Weise et al., 2020, Van Deventer et al., 2020, Nhamo et al., 2017, Mueller et al., 2016). In most cases, datasets are made open access with collaborators including UN organizations, the International Union for the Conservation of Nature (IUCN), the EC-JRC, the World Resources Institute (WRI), the German Aerospace Centre (GAC), NASA, and EU Horizon 2020. Of particular note are the advancements made by Pekel et al. (2016), which have supported the development of the Global Surface Water Explorer which is used by the UN Environment Programme (UNEP) for global monitoring of indicator 6.6.1 and is also now being tested by countries (**Box 3-5**).

7 AFFORDABLE AND CLEAN ENERGY

Studies of relevance for monitoring **SDG 7** on energy, include the use of earth observation and crowdsourced OSM data, geospatial modelling and machine learning methods to map access to electricity (7.1.1) (Arderne et al., 2020, Stokes and Seto, 2019, Falchetta et al., 2019), global wind and solar installations using OSM and machine learning (7.2.1) (Dunnett et al., 2020), and modelling of least-cost electrification options using OSM and satellite data in sub-Saharan Africa (7.2.1) (Mentis et al., 2017). In most cases, the datasets and source code for these studies are made open access. Collaborators include the World Bank, IIASA, NASA, the EC-JRC, and the UN. Of particular note are advancements in estimating electricity access using satellite imagery, which have inspired a new tool developed by the World Bank to support national monitoring of electricity access (**Box 3-6**).

8 DECENT WORK AND ECONOMIC GROWTH

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

10 REDUCED INEQUALITIES

Studies relating to **SDGs 8** (economy), **9** (infrastructure) and **10** (inequalities) were comparatively limited. Wang et al. (2019) use machine learning and NTL data to develop a gridded global map of gross domestic product (GDP) as well as national Gini indexes (8.1.1, 10.4.2), while Ivan et al. (2020) use NTL and other satellite and administrative data to produce an NTL Inequality Index for Romania (10.4.2). In relation to SDG 9, Oshri et al. (2018) use satellite imagery and survey data with deep learning methods to produce a map of infrastructure (electricity, sewerage, piped water, and roads) for Africa (9.1.1). Datasets are not made open access, with some limitations due to the confidentiality of household survey data inputs.

11 SUSTAINABLE CITIES AND COMMUNITIES

A comparatively large number of studies (12 studies) corresponded to **SDG 11** on cities. Of relevance for indicator 11.1.1, two papers apply deep learning methods and high-resolution satellite imagery to identify the extent of urban slums in selected major cities (Stark et al., 2020, Wurm et al., 2019), while another study maps unimproved housing in Africa using survey data and geostatistical regression modelling (Tusting et al., 2019). Only the latter study provides the dataset and all code open access. Collaborators include the GAC, UN organizations, the BMGF, and EU Horizon 2020.

For 11.3.1 on land-use efficiency, Schiavina et al. (2019) use the Global Human Settlement Layer (GHSL) to estimate land-use efficiency at the global, regional, and national levels, while Zhou et al. (2018) develop spatially and temporally consistent global urban maps from 1992 to 2013 using NTL data. These global datasets are not made open access. A recent study develops a new open access and very high-resolution (10m) global gridded human settlement map for the year 2015 (World Settlement Footprint) using machine learning and optical (Landsat-8) and radar (Sentinel-1) satellite imagery (Marconcini et al., 2020). Collaborators include the EC-JRC, the GAC, Google, the European Space Agency (ESA), MindEarth, and NASA. The high-resolution World Settlement Footprint has attracted considerable interest from several international organizations who are now using the dataset as well as new emerging products for a range of applications, particularly in urban planning and disaster response (**Box 3-7**).

Three further studies model particulate matter (PM2.5, PM10) concentrations and population exposure relating to 11.6.2 at global scales, including in the context of the GBD assessments which are made open access (Shaddick et al., 2020, Shaddick et al., 2018) and for European countries (Beloconi et al., 2018) using Bayesian geostatistical modelling approaches. An additional two studies test methods for monitoring green open space in cities (11.7.1) including a global study (Corbane et al., 2020), however, datasets are not made open access. Collaborators include NASA, the WHO, and the EC-JRC.

Studies on **SDG 13** on climate action were more limited and included the use of big data to generate real-time estimates of reduction in CO2 emissions (13.2.2) in 2020 as a result of the COVID-19 pandemic (Le Quéré et al., 2020), as well as new high-resolution and open access spatial datasets (TerraClimate, CHIRPS) on essential climate variables (Abatzoglou et al., 2018, Funk et al., 2015). Collaborators include EU Horizon 2020, the Gordon and Betty Moore Foundation (GBMF), the USGS, NASA, and Google.

Of relevance for 14.1.1(b) on ocean debris, Biermann et al. (2020) develop and test an automated method for the detection and classification of floating plastic materials from Sentinel-2 multispectral imagery using a Naïve Bayes algorithm trained to predict plastics. Other studies of relevance for **SDG 14** include the use of satellite imagery and machine learning to map the global distribution and trajectory of tidal flats (Murray et al., 2019) and a composite dataset of the global extent of salt marshes (Mcowen et al., 2017). While not specifically addressed in the SDG indicators, these datasets would be relevant for monitoring the sustainable management of coastal areas as per SDG targets 14.2 and 14.5. Both datasets are provided open access. Collaborators include UN organizations, the Global Environment Facility (GEF), the Nature Conservancy (TNC), and Google.

The largest number of papers reviewed (21 studies) corresponded to SDG 15 (life on land). Studies relating to 15.1.1 provide methods for evaluating forest degradation (Mondal et al., 2020) and disturbance (Hansen et al., 2016) combining satellite imagery (Landsat, Sentinel, MODIS) with machine learning algorithms (classification trees, random forest). Another study by Mondal et al. (2019) evaluates the use of random forest and regression trees with Sentinel-2 data and the Google Earth Engine (GEE) cloud computing platform to classify mangrove forests. A recent study by Hansen et al. (2019) develops two data products, the Forest Structural Condition Index (SCI) and the Forest Structural Integrity Index (FSII). Of these, only Hansen et al. (2016) and Hansen et al. (2019) make their datasets open access. Two further studies map above-ground biomass using satellite imagery and spatial datasets (Forkuor et al., 2020, Avitabile et al., 2016) of relevance for 15.2.1 on sustainable forest management as well as REDD+ reporting. Collaborators include the WRI, NASA, the UN, Google, the Wildlife Conservation Society (WCS), the ESA, the GBMF, and several bilateral donors.

Studies relating to 15.1.2 on the protection of important sites for terrestrial biodiversity develop new global-scale maps for terrestrial world ecosystems (Sayre et al., 2020), natural and modified habitat (Gosling et al., 2020), and protected area connectivity (Saura et al., 2019). Several studies also focused on monitoring 15.3.1 on land degradation at various scales, including national studies (Wang et al., 2020, Mitri et al., 2019) and global-scale studies on mapping land-use change (Song et al., 2018), soil organic carbon (Hengl et al., 2017) and land degradation (Giuliani et al., 2020) (**Box 3-8**). Other studies use satellite imagery as well as machine learning methods to develop global datasets relating to 15.4.2 on mountain green cover index (MGCI) (Bian et al., 2020), 15.5.1 on threatened species (Jung et al., 2020), modelling of invasive species of some relevance for 15.8.1 (Jensen et al., 2020, Ouko et al., 2020), as well as text mining of national commitments relating to agrobiodiversity of some relevance for 15.9.1 (Juventia et al., 2020). Collaborators include the USGS, Esri, TNC, UN organizations, the EC-JRC, ESA, the GEF, NASA, the GBMF, WRI, IIASA, BirdLife International, NatureMap, and Norway's International Climate and Forest Initiative (ICFI).

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

Finally, studies of relevance for **SDG 16** on peace and security include health and mortality-related indicators addressed through the GBD studies (16.1.1, 16.1.2, 16.1.3, 16.2.3) (GBD 2016 SDG Collaborators, 2017). Li et al. (2019) develop a method using machine learning and Twitter data to detect reports of corruption (16.5.1), while Levin et al. (2018) examined the potential of NTL data and mining of news events and Flickr photos for monitoring crisis development and refugee flows. Both studies could be considered exploratory testing of methods, and datasets are not made available.

TRENDS
Thematic Research Network
on Data and Statistics

Partners
for
Review