COLLABORATIVE ON SDG DATA INTEROPERABILITY

Global Partnership for Sustainable Development Data

USING DATA TO JOIN UP DEVELOPMENT EFFORTS

WHY **INTEROPERABILITY?**

The 2030 Agenda for Sustainable Development comprises 17 Sustainable Development Goals (SDGs) made up of 169 targets and 232 indicators. In addition to their broad scope, targets are multidimensional; covering socio-economic, environmental, and inequality-related issues. Moreover, given the call to leave no one behind, data for the SDGs need to be produced to varying degrees of disaggregation. This includes a geospatial focus given the ongoing need for planning and progress tracking at sub-national levels. There are numerous opportunities to plug information gaps that exist both across geographies and sectors; for example, harnessing the potential of big data for development or Earth observation data. However, these opportunities can only be effectively leveraged if the data that is collected and processed is produced in formats and organized in ways that are 'interoperable'. This involves multiple stakeholders agreeing to follow common principles and procedures that allow for data standardization, comparability, and integration.

As development professionals, we must consider how we can optimize the data standards we use in our sector to be more interoperable. For instance, take this simple analogy. When sending emails between Macs, PCs, and smartphones and tablets of any brand, a specific internet protocol embedded in the world wide web ensures that the email is transmitted in a way that enables a user at the receiving end to read it, regardless of what operating system or hardware they are using. This is an example of a standardized approach to interoperability that enables systems to talk to each other. The email itself may contain attachments – documents, spreadsheets, photographs, etc. – that exist in many different formats. Whether we are using a Mac or PC, are at a desktop or on a smartphone, we will in all likelihood still be able to access the attachments. This too is because the 'data standards' that underpin programmes like Outlook, Apple Mail, Gmail, and others have been designed to be interoperable despite the fact that they are sold by different companies. This is because despite their competitive nature, there is recognition that the development of interoperable technologies can confer such enormous benefits for users that it can benefit the technology industry as a whole. (Palfrey et al 2012).

Similarly, in the sustainable development field, there is an unrealized opportunity to extract far more value from the vast quantities of data that exist – whether they be in Aid Information Management Systems (AIMS), surveys, censuses, satellite images, citizen-generated data, or any other source. Investing time and resources in the development and deployment of sector-wide interoperability solutions offers us all collectively an opportunity to make far more use of the data that currently sits in sectoral and institutional silos. Such solutions will simplify our ability to access, share, manipulate, and use complex data with as little effort as we currently exert to access emails and their attachments.

Before we can reach this end goal however, we must ensure we have the information needed to achieve and measure the targets of the 2030 Agenda for Sustainable Development (2030 Agenda). To do so, government bodies and development professionals must first better understand the role that interoperability must play. In particular, it is vital that operational staff within government ministries, departments and agencies (MDAs), not just national statistical offices (NSOs) or IT staff within MDAs, recognize and implement the issues covered in this note. This is important so that they can make the right decisions about how data is collected, made available and used, and so that integration and comparability challenges are not left until after information systems have already been produced, developed, and are in place. It is also important to help avoid duplication of efforts and the prevention of the same, or very similar, data being collected multiple times by different institutions because there is little or no coordination between them.

This document is an introductory note to “Interoperability: A practitioner’s guide to joining-up data in the development sector” (the Guide). Its purpose is to introduce the concepts discussed in the Guide and to highlight how a systemic approach to handling data interoperability issues can support both the achievement and monitoring of the SDGs. The primary audiences for this note, and for the Guide, are official statisticians, development practitioners responsible for data management, and the suppliers of ICT solutions and services in the development sector.

INTEROPERABILITY SOLUTIONS: **WHAT CHALLENGES CAN INTEROPERABILITY SOLVE?**

At the moment, data are under-utilized, representing significant untapped potential in all the data that are available within national data ecosystems. This is true regardless of where data are sourced from (surveys, administrative systems, Earth observations and sensors, etc.), or existing levels of ability and capacity. In many countries, this situation results in a systemic failure to take advantage of possible new sources of information to help fill knowledge gaps and guide policy decisions. The ability to easily and efficiently share data is crucial to overcoming these challenges.

> *"Interoperability is the ability to join up data from different sources in a standardised and contextualised way. However, it is about more than just the form and structure of data, it is also about solving problems in a joined-up way. [...] Interoperability can help reduce the time, effort and expense involved in data collection; eliminate the frustration and risks associated with handling inconsistent and incomplete data; and meet the need for internationally comparable, sustainable, disaggregated data to ensure that no one is left behind." (JUDS 2016, 1 - 7).*

When thinking about the use of data at a systemic scale, to make information more shareable, data, the building blocks of information, need to be made more shareable to enable widespread reusability. Approaching the problem on a 'systemic scale' means exploring how data sets, IT systems, internal data sharing and management, and cross-organizational coordination around data issues can be better organized and structured, with appropriate protocols and policies put in place across government and sectors.

> *To scale-up the use of data, information needs to be provided about the data ('metadata') being produced, published, and shared to make it comprehensible to others who will also use it.*

This process of systematizing data is very context-specific and reliant on numerous factors. Some factors can be technological, relating to the physical machinery needed to share data. Other factors can be digital, relating to whether or not data that once existed in paper documents and charts has been converted into an electronic form, and if so, whether it is being stored in formats that are comprehensible to both humans and machines. Yet other factors relate to how people and whole organizations manage, organize, and coordinate the governance of their data sets and systems. The Data Commons Framework (Goldstein et al 2018) can help us break down and logically understand how these various dimensions relate to one another.

The 2030 Agenda encompasses socio-economic development, environmental protection, and tackling economic inequalities on a global scale. Keeping this broad scope in mind, the Data Commons Framework (see Figure 1) can help us better understand the 2030 Agenda's data interoperability needs; particularly within the three broad semantics layers: organizational practices; institutions, law and policy; and, humans.

Figure 1: Data Commons Framework (Goldstein et al 2018)

> *Sharing data and information across the professions and sectors of the sustainable development field necessitates the professionals working in the field to have a common understanding of the 'semantics' used by other branches of the field. In other words, having a common understanding of what particular terms mean.*

For example, to understand the impact of climate change on macro-economic trends, development economists must learn and understand climate science terms. Similarly, as the fields of statistics and data science edge closer together, statisticians are having to learn whole new vocabularies and concepts that will help them disseminate and share their products in new ways. For instance, ensuring that statistical data can be presented online on interactive maps that integrate it together with data obtained from satellites and other observational and sensory sources.

New ontologies – ways of organizing vocabularies and concepts which contain the definitions for particular terms – are needed to help rationalize the exchange of knowledge between professions both on a human and machine level.

THE NEED FOR **COORDINATION, COLLABORATION AND GOVERNANCE**

Because the sustainable development field is global – the 'indivisible', 'holistic' and 'universal' dimensions of the 2030 Agenda are its core attributes – it is not enough for individual professionals to have a common understanding of the language of sustainable development. Government MDAs, NSOs, intergovernmental organizations including UN agencies, funds, and programmes, NGOs, and other interest groups all must be able to interpret and share information in a way that is logical and useful.

> *'Data governance', how data is controlled, and 'data management', how data is processed and handled, must become integral components of organizational strategies and business processes.*

In this way, whole organizations and sectors can transform how they use their data, becoming 'data-driven' in their operations and outlook. This requires coordination, organizational oversight, and accountability to implement changes to long-held institutional behaviors that have historically tended to pigeon-hole data issues to the IT department, rather than treating them as cross-cutting issues that require an organizational response, effective governance, and management.

For example, much of the data needed to achieve and measure the SDGs can be sourced from administrative data sets across the national statistical system, not just from NSOs. To better leverage this opportunity, it is necessary to develop interoperability frameworks to ensure there is consistency in how data is shared and produced between NSOs and other MDAs that are the custodians of administrative data sets. It is also necessary to foster inter-institutional collaborations to proactively promote collaborative, inclusive approaches to the design of data sets and processes.

> *All of us - from sector specialists and development professionals, to IT experts and end-users-need to recognize the imperative for interoperability.*

Being able to share data and ensuring that professionals, organizations, and even computers themselves, have a common understanding of what different terms mean, is an enormous challenge. However, it is central to coordination and measurement efforts to achieve the SDGs

> *Despite the complexity and range of the issues at play, one key characteristic of good quality data stands out as a concept that binds everything together and holds potential as a framework for helping overcome them. The characteristic is 'data interoperability'; the ability to join-up data without losing meaning.*

One of the unique properties of data is the value that emanates from its reusability as a resource. Data can be reused in many different ways by different entities and people. It is because of this reusability that today's most valuable listed companies collect data for resale and reuse in processing and analysis.

In the sustainable development field, the types of data that are commonly collected – for instance demographic data, data on the condition of the natural environment, or human migration, among many others – may not have the same short-term financial value as the personal data that is harvested by companies such as Google and Facebook, but they hold enormous value to policymakers and development professionals tasked with finding ways of improving socio-economic and environmental conditions. Ensuring that the data collected is interoperable from the outset is crucial to making sure that it will later be shareable in a useful and understandable way and, by extension, more reusable and therefore valuable.

THE ROLE AND WORK OF THE **COLLABORATIVE ON SDG DATA INTEROPERABILITY**

The Collaborative on SDG Data Interoperability was formed in 2017 following the first UN World Data Forum (UN-WDF) hosted by Statistics South Africa in Cape Town, South Africa. The Collaborative's members recognize the value of interoperability as a framework that can be used to improve how data flows across the sectors and professions of the sustainable development field. They recognize that in this way, interoperable data can contribute to improved outcomes, achievement of the SDGs and more efficient and accurate monitoring of development targets. Following extensive consultation at the Data for Development Festival in Bristol in March 2018, the Collaborative agreed to focus on the production of practical guidance on data interoperability for development professionals.

Interoperability: A practitioner's guide to joining-up data in the development sector is the Collaborative's attempt to produce guidance designed for statisticians, development professionals who manage data, and suppliers of ICT solutions with a common reference point for interoperability issues. The Guide is designed to be a living document that will both inform and be informed by common practices in the development sector. Its contents identify and cover five cross-cutting areas, or 'dimensions', of data interoperability that capture the broad range of issues and challenges that Collaborative members have highlighted as being integral to solving in order to make data interoperable at a systemic scale within the sustainable development field.

The five areas covered by the Guide include:

1. DATA MANAGEMENT, GOVERNANCE AND INTEROPERABILITY

Broadly speaking, the technologies and methods needed to make data speak to each other already exist. The most serious impediments to interoperability often relate to how data is managed and how the lifecycle of data within and across organizations is governed. Data management, "the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles" (DAMA 2017, 17) is therefore the cornerstone of any effort to make data more interoperable and reusable on a systemic scale. To be effective, data management requires that data be governed – well controlled with oversight and accountability – across its lifecycle.

This section of the Guide explores the concepts of data interoperability and integration, management and governance in more detail; setting out how interoperability can be used as a framework to improve data management and decision-making, its relationship to the SDGs, and highlighting some useful institutional tools and examples that can help practitioners in the development of their data management and governance strategies. At its heart, this section extols the benefits of thoughtful planning, continuous strategic management and governance of data across its lifecycle, and consideration of user needs from the outset when striving to modernize IT systems and amplify the reusability and audiences of existing data.

2. CANONICAL DATA AND METADATA MODELS

A significant data interoperability challenge in the development sector relates to how the structure and description of data and metadata – data about data such as the author/producer of the data set and the date the data was produced – can be organized consistently. Having standardized or interchangeable models is especially important for interoperability as this is what enables computers to automatically identify data and provide users with information on where the data originated, when it was last updated, etc. In other words, interoperability is highly dependent on data and metadata modelling decisions and practices.

As it currently stands, different organizations, and even different departments within organizations, often approach data and metadata modelling on a case-by-case basis, adopting lots of different approaches that are not designed with broader data-sharing in mind. Such models usually prioritize internal needs over the needs of broader user groups. One interoperability-specific challenge emerges from the fact that there is usually no single "right" way of representing information, with some data structures being better suited for managing transactional processes (e.g., capturing data from a survey or maintaining a civil registration database) and others being better suited for analyzing and communicating data to users (e.g., for the creation of data visualizations in a monitoring dashboard). This challenge is compounded by the fact that people often model data in isolation with a specific application in mind. As a result, the same information content is often represented in variety of (usually incompatible) ways across different systems and organizations. Canonical models refer to data models that follow specific standardized patterns that make them highly reusable and conducive to data sharing. For example, if different applications are pulling data on the same data set, each application must independently understand and transform different data set structures. But canonical models can be used to represent multiple sources of data and metadata using common patterns, thus making data integration simpler and more efficient.

3. CLASSIFICATIONS AND VOCABULARIES

The third dimension of data interoperability that the Guide explores relates to data classification systems. Classification systems shape the way data is collected, processed, analyzed and shared with users. They constitute the basis for data management and data interoperability. The use of common classifications and vocabularies allow data to be shared efficiently and for users to more easily find related information across numerous data platforms.

Members of different data communities are increasingly collaborating in the development and use of common classifications and vocabularies to describe cross-cutting concepts and relationships across different information sources and professional sectors. For example, the Committee on Data (CODATA) of the International Council for Science is undertaking efforts to generate common classifications and vocabularies across science disciplines and PEPFAR is supporting efforts to join-up classifications used by a range of entities across government and professional disciplines that work on the HIV/AIDS epidemic in various ways.

Systemic data interoperability requires that common classifications and standard vocabularies across different applications or software systems are used. To remain effective, such classifications and vocabularies need to evolve over time and be 'mapped' – the process of programming a computer to automatically understand the relationships between the concepts and terms used to label data – to each other.

4. STANDARDIZED INTERFACES

As computers' processing power and internet speeds have increased, it has become a lot easier to present data online in far more visually exciting and engaging ways. A user's experience of how they interact with data on a website has become a key indicator of a platform's quality. Graphics, visualizations, and other often interactive tools contribute to an enhanced user experience. Using the world wide web as a conduit, applications can be programmed to share data so that the same data can be integrated with information hosted on other platforms and presented in numerous different ways depending on user need. This process requires the use of Application Programming Interfaces (APIs) – virtual highways that allow data to travel back and forth between different websites and platforms.

Adopting and implementing web-based APIs that follow common standards and well documented patterns enables multiple developers to produce interactive data-driven applications. It also creates new possibilities for user engagement and on-the-fly data analysis and visualization. Similarly, using standardized patterns and reusable building blocks when designing human-machine interfaces across different applications can significantly reduce the effort that users need to invest in order to find and use the data they need. High-quality APIs should be built with application developers' needs in mind, focusing on helping them create applications that satisfy end users' needs.

5. IMPLEMENTATION OF LINKED-DATA APPROACHES

In many ways this final dimension builds on most, if not all, previous parts of the Guide. The Internet has created unprecedented opportunities for the open exchange of information and the creation of knowledge. When Sir Tim Berners Lee created the world wide web as a human-readable interface for the sharing of information over the Internet, he also envisaged that a 'semantic web' – a machine equivalent of the world wide web that would enable computers to automatically read and understand data on the Internet – may emerge. While the semantic web has not materialized (yet) at scale, linked-data approaches that underpin the semantic web concept hold a lot of prospective value for the development sector.

The process of information discovery and knowledge creation is significantly enhanced by the ability to establish meaningful links between independently produced and managed information resources. This is particularly important in the context of data for development, as the 'indivisible' nature of the SDGs makes it more urgent than ever to join-up the information resources and data assets owned and managed by different sectors and communities (for instance, linking administrative data on financial expenditure by local governments to performance indicators of poverty-reduction programs and official unemployment statistics). There is growing interest in tools and technologies that allow for the publication of data in such a way that users can easily identify and integrate semantically related information resources over the web. In particular, data managers are increasingly interested in implementing principles of "linked open data" in the dissemination of their information assets.

NEXT STEPS:
SCALING UP TO SYSTEMIC APPROACHES

When taken together as a whole, these five dimensions of interoperability cover most of the areas needed to help scale interoperability solutions to systemic levels. This is intentional. The Guide has been developed as an iterative and practical tool, to be used by development professionals across interest groups to help improve the reusability of their data and data systems by making them interoperable. It is hoped that new sections, examples, and guidance will be added over time to ensure its continued relevance and usefulness in this fast-evolving space.

The Guide's chapters are designed to reflect its practicality. Each chapter offers an overview of key messages and themes, and defines key concepts before delving deeper into particular resources and examples. Each chapter also concludes with a brief literature review and a Building a roadmap section. This is a key component of the Guide's practical application. When taken together, these Roadmap components will set out an assessment framework that data managers in development organizations can use to assess the degree to which their systems are interoperable or not, and where further action is required. As with the Guide in general, it is hoped that this assessment tool is developed further in coming years and applied by organizations across stakeholder groups.

The Collaborative on SDG Data Interoperability will continue to build and maintain the document as it develops as a tool. It is hoped that new synergies will form between data producers, publishers, users, and those providing capacity building and training. In this way, the guidance set out within the Guide can be incorporated into existing training materials and modules and we, and our computers and data systems, can all start speaking the same language.

COLLABORATIVE ON SDG DATA INTEROPERABILITY

Global Partnership for Sustainable Development Data

References

DAMA International. (2017). Data Management Body of Knowledge (2nd Ed). New Jersey: Technics Publications.

Goldstein, E., Gasser, U. & Budish, R. 2018. Data Commons Version 1.0: A Framework to Build Toward AI for Good [online], available at: https://medium.com/berkman-klein-center/data-commons-version-1-0-a-framework-to-build-toward-ai-for-good-73414d7e72be [Accessed 2 October 2018].

Joined-Up Data Standards project. (2016). The frontiers of data interoperability for sustainable development, [online]. Available at: http://devinit.org/wp-content/uploads/2018/02/The-frontiers-of-data-interoperability-for-sustainable-development.pdf [Accessed 13 September 2018].

Palfrey, J. & Gasser, U. (2012). Interop: The promise and perils of highly interconnected systems. New York: Basic Books.

Contact us

This note has been produced on behalf of the Collaborative on SDG Data Interoperability, convened by the UN Statistics Division (UNSD) and the Global Partnership for Sustainable Development Data (GPSDD). For further information, please contact info@data4sdgs.org

Acknowledgements

Special thanks to the members of the Collaborative on SDG Data Interoperability, the UN Statistics Division, the Global Partnership for Sustainable Development Data secretariat, and Open Data Watch.


COLLABORATIVE
ON SDG DATA
INTEROPERABILITY


Global
Partnership
for Sustainable
Development Data