Statistical Commission
Fiftieth session
5 – 8 March 2019
Item 3(c) of the provisional agenda
Items for discussion and decision: Open data

Background document
Available in English only

A review of open data practices in official statistics and their correspondence to the Fundamental Principles of Official Statistics

Prepared by the United Nations Statistics Division

A review of open data practices in official statistics and their correspondence to the *Fundamental Principles of Official Statistics*.

A. Introduction

Because of the centrality of statistics in setting policies and measuring their outcomes, national statistical offices (NSOs) and national statistical systems (NSSs) should be at the forefront of the data revolution and the open data agenda. Although their specific responsibilities differ from country to country, NSOs generally have the authority to set statistical standards; to design and implement large-scale data collection programs; and to ensure the quality, reliability, and availability of official statistics. NSOs generally also have the trust of citizens and governments to make data open without breaking privacy or confidentiality. By collaborating with other NSOs and international statistical agencies, they contribute to and benefit from technological innovation, the development of new methodologies, and the adoption of common standards. For NSOs and their partner agencies across NSSs, open data (see Box 1) is more than a dissemination strategy: embracing the principles of open data is an opportunity to engage with the larger world of data-driven innovation, potentially leading to economic value[1], cost savings, and process improvement and to demonstrate their relevance to their own governments, the private sector, and the public at large.

The United Nations Statistical Commission's Friends of the Chair Group on the Implementation of the Fundamental Principles of Official Statistics (FOC-FPOS) has undertaken a comparison between the 10 United Nations Fundamental Principles of Official Statistics (FPOS) and the 6 Principles of the Open Data Charter (ODC). The results of this comparison can be found in the Appendix B.

Set in the context of the above comparison, this background document firstly examines the practical application of open data principles in official statistics, with a focus on challenges related to open data standards; data interoperability; public engagement; and protection of data privacy. The paper then discusses the capabilities and activities necessary to deliver open data. The third section looks at the emerging issues that arose through the recent review of the Open Data Charter, namely: openness by default, data sovereignty, data governance, management and infrastructure, and data privacy, security, and confidentiality. Finally, the paper draws all this together in its conclusion.

B. Practical application of open data in official statistics

This section will explore some of the key issues that emerge when NSOs seek to make their statistics and datasets available as open data: open data principles; data interoperability; licensing; public engagement; and protection of data privacy. This section provides a broad introduction to these issues and points to further resources on the subject.

Implementing open data principles

Implementing open data means operationalizing open data principles, such as the requirements of the Open Definition or the principles of the Open Data Charter. Guidance on the implementation of the Open Definition is provided in OKI's *Open Data Handbook*. Additional materials on the components, dimensions, and applications of open data can be found from a variety of sources, including: the Open Data Institute's (ODI) series of *Guides,* the World Wide Web Foundation's (Web Foundation) *Research* section, and materials from the Open Data Charter's Resource Centre. All these materials, and many more,

[1] https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf

are freely available for use by NSOs seeking to better understand the opportunities available for them to leverage the benefits of open data for official statistics.

Box 1. Definition of Open Data.

Open Knowledge International's (OKI) *Open Definition* [2] provides a short, simple definition of open data:

> *Open data is data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike.*

The Open Definition 2.1 states four requirements for open data:

1.1 **Open License or Status** - The work must be in the public domain or provided under an open license.
1.2 **Access** - The work must be provided as a whole and at no more than a reasonable one-time reproduction cost and should be downloadable via the Internet without charge.
1.3 **Machine Readability** - The work must be provided in a form readily processable by a computer and where the individual elements of the work can be easily accessed and modified.
1.4 **Open Format** - The work must be provided in an open format. An open format is one which places no restrictions, monetary or otherwise, upon its use and can be fully processed with at least one free/libre/open-source software tool.

The Open Data Charter follows a similar schema, with four principles that define openness: 1) Open by Default; 2) Timely and Comprehensive; 3) Accessible and Usable; 4) Comparable and Interoperable. And two that describe the purpose of open data: 5) For Improved Governance and Citizen Engagement and 6) For Inclusive Development and Innovation.

Open Data Watch (ODW) has operationalized the Open Definition in its Open Data Inventory (ODIN) methodology, which assesses data coverage and openness of national statistical systems. The ODW Openness Assessment has five elements, namely: (1) machine readability; (2) use of non-proprietary formats; (3) availability of multiple download options; (4) availability of metadata providing sufficient context to understand the data; and (5) open licensing. ODW's assessment methodology is available to NSOs or other statistical agencies for self-assessment.[3] Countries wishing to improve the openness of their data can do so at a relatively low-cost by: providing data in machine-readable formats; making metadata available; and publishing open terms of use. Without machine readability, perhaps the most important of the elements, users cannot easily access and modify the data, which severely restricts the scope of the data's use. More information on this and the other elements in ODIN are dealt with in more detail in Appendix A.

[2] http://opendefinition.org/
[3] http://odin.opendatawatch.com/Downloads/otherFiles/ODIN-2017-Methodology.pdf

Interoperability

Data interoperability is the ability to easily extract data and to use it and integrate it with other datasets across different systems[4]. It is therefore an enabler of open data and a pre-condition for data to have impact on policy and decision-making process. Moreover, data interoperability is a multi-dimensional characteristic of good quality data, which requires adequate institutional and governance frameworks; the adoption of standard data and metadata models, classifications and vocabularies for structuring and describing information; and the use of standard technological platforms, interfaces and protocols to allow users to find, link, and integrate datasets from different sources, both manually and by automated means, into their own applications. The Collaborative on SDG Data Interoperability[5] convened by the UN Statistics Division and Global Partnership for Sustainable Development Data (GPSDD) launched *Data Interoperability: A practitioner's guide to joining up data in the development sector*[6] at the 2018 World Data Forum, held in Dubai, UAE in October 2018. The Guide identifies five dimensions of interoperability that are required for the development of data systems and processes capable of integrating data from numerous sources, including data published in open data-friendly formats.

The implementation of interoperability standards for the publication of open data by NSOs and other organizations also requires the adoption of common metadata schemas, vocabularies and classifications to describe individual datasets. For example, the Data Catalogue Vocabulary (*DCAT),* and its many derivatives, is recommended by the *World Wide Web Consortium* (W3C) — an international community of experts who develop web-based standards — as standard for structuring metadata in an open and interoperable way.

Since the 1990s, the accelerated development of Web technologies has made the work of finding, merging, and linking data across systems much easier. Modern tools for data exchange and dissemination on the web, such as web-services based on open APIs, can be used by multiple users to run their different analytic applications with the most up-to-date information as soon as the data becomes available. Further, standardized interfaces and bulk downloading options can make it much easier for users to find and access data over the web, and to seamlessly integrate it into their own business processes. Taking advantage of new technologies for automation of data integration (e.g., through Artificial Intelligence) is now a priority for statistical organizations, particularly in the face of the increasing amount of data that is generated across society and needs to be collected, processed, analysed and openly disseminated to support informed decision-making at all levels.

However, the implementation of technical standards and solutions to improve data interoperability in the context of legacy systems and architectures (which are often characterized by unknown dependencies and incomplete documentation) requires difficult, complex, and often costly organizational and behavioural changes. This includes the establishment of new processes and governance mechanisms, as well as the investment of resources to develop new skills and build capacity to implement new standards, technologies and tools.

Furthermore, to fully maximize the benefits of open data, it is important that statistical organizations prioritize data interoperability standards which are commonly used by a broad range of stakeholders. Although data interoperability standards are a deeply technical issue, their practical implementation is still

---

[4]Liz Steele and Tom Orrell, 2017, *The frontiers of data interoperability for sustainable development.* Available at: http://www.publishwhatyoufund.org/wp-content/uploads/2017/11/JUDS_Report_Web_061117.pdf

[5] http://www.data4sdgs.org/initiatives/interoperability-data-collaborative

[6]Luis Gonzalez Morales and Tom Orrell, 2018, *Data Interoperability: A practitioner's guide to joining up data in the development sector.* Available at: http://www.data4sdgs.org/sites/default/files/services_files/Interoperability%20-%20A%20practitioner's%20guide%20to%20joining-up%20data%20in%20the%20development%20sector.pdf

highly variable between countries, and should be driven by the needs of users beyond specific local and national contexts.

Public engagement

As the data ecosystem expands, NSOs are expected to take a stronger coordinating role encompassing new data sources, producers, and users, including both public and private actors. NSOs must now engage with an increasingly diverse set of stakeholders, including government agencies, academic institutions, non-governmental organizations (NGOs), businesses, and bilateral and multilateral institutions. To adopt the "leave no one behind" principle of the 2030 Agenda for Sustainable Development, NSOs need to build a broad coalition of all segments of society and make sure all producers and users of data are counted and benefit from the systematic implementation of open data principles across the NSS. By embracing open data principles and practices, NSO can raise their standing as the trusted institution that ensures all users have ready access to high-quality data and statistics that meet national and international demand for information, while protecting privacy and confidentiality in line with the Fundamental Principles of Official Statistics. In embracing open data principles and practices, there is a responsibility on the NSO to adhere with agreed standards and best practices.

NSOs should be, by their design, apolitical government organizations. Politics, however, can become entangled in NSO activities as official statistics are often used to justify funding decisions from donors[7] or governments,[8] including fiscal policy and other functions of state power.[9] In this context, NSO leadership often lack or are hesitant to use their political capacity to push for an open data agenda.[10] There is need for a national consensus and high-level commitment by governments to support a long-term open data movement, providing the necessary political backing to introduce necessary changes in national data policies and infrastructure. Therefore, instead of focussing only on the technical challenges of producing data and statistics, NSOs should also invest effort into documenting successful applications that demonstrate the value of high-quality, trusted, and open data for policy and decision making at all levels, with a view to increase support for open data policies across the NSS.

It is important that NSOs undertake a consultation with their local (prospective) user groups before embarking on a program to open their data. Every national and subnational context is unique and, to the extent possible, data users should be consulted on how their needs could be met. This user-centred approach can help to build trust in the NSS as well as enable the emergence of new innovations and business models that rely on open statistical data.

An important first step is to secure political and institutional support for open data in official statistics within the government and obtain the support of other stakeholders. This effort should be coordinated with any existing government-wide open data initiative. The legal framework and access-to-information

---

[7] Justin Sandefur and Amanda Glassman, 2014, *The Political Economy of Bad Data: Evidence from African Survey & Administrative Statistics.* Available at: https://www.cgdev.org/publication/political-economy-bad-data-evidence-african-survey-administrative-statistics-working.

[8] Samantha Custer. and Tanya Sethi (Eds.), 2017, *Avoiding Data Graveyards: Insights from Data Producers and Users in Three Countries.* Available at: http://docs.aiddata.org/reports/avoiding-data-graveyards-report.html.

[9] Florian Krätke and Bruce Byiers, 2014, *Implications for the Data Revolution in Sub-Saharan Africa.* Available at: http://ecdpm.org/wp-content/uploads/DP-170-Political-Economy-Official-Statistics-Africa-December-2014.pdf.

[10] World Bank, 2017, *World Bank support to open data 2012-2017*. Available at: http://opendatatoolkit.worldbank.org/docs/world-bank-open-data-support.pdf.

policies should be reviewed and revised as necessary to support open data policies. Open data should be incorporated in countries' National Strategies for the Development of Statistics (NSDS) – as Ghana has done with their 2017-2021 NSDS[11] – as well as in the planning and implementation of SDG national reporting platforms. Countries can also carry out an Open Data Readiness Assessment (ODRA)[12] as the basis to identify a road map for implementing a national open data policy. And, just as NSOs should champion open data in their own countries, their perspectives and voices are needed at international discussions around open data such as the International Open Data Conference and United Nations World Data Forum.

NSOs should also consider participating in (or establishing, where none already exists) domestic, regional, and international multi-stakeholder networks that bring official and non-official data producers and users together to coordinate, explore, and improve data systems. Here are some international networks that NSOs should consider engaging with: the *Open Data for Development Network* (OD4D), *Open Government Partnership* (OGP), *Global Partnership for Sustainable Development Data* (GPSDD), *Global Open Data for Agriculture and Nutrition* (GODAN), and Open Data Charter. Not all of these networks will be appropriate for all NSOs, however, collectively they offer a way to reach out to open data user groups and provide avenues for staying up-to-date on new open data practices and innovations. Beyond coordination efforts and building political support, NSOs can engage the public through their websites and open data portals.

To facilitate reporting and public engagement, the UNECE has created a practical guide on national reporting platforms for the SDGs.[13] The development by Open Data for Development of regional hubs[14] that support open data as well as the inclusion of NSO representatives at the IODC and UNWDF[15] are also important developments that bring more engagement between the open data and official statistics communities.

Data Privacy

National statistical systems are the repositories of two kinds of data: microdata — which are the unit records of censuses, surveys, and administrative datasets — and aggregate statistics compiled from microdata. Raw microdata contains individually identifiable information about people, businesses, or other entities. Therefore, before microdata can be made disseminated, they must be anonymized or aggregated into data files suitable for public or licensed use using tools such as SDCMicro[16]. Access to the underlying microdata must be strictly controlled using various accountability mechanisms, such as requiring users to register, to agree to strict terms of use, and describe exactly who will use the data, and how they will use it. Some countries only allow microdata downloads after a rigorous case-by-case review process. Accordingly, countries must find a balance between protecting respondent information from potentially malicious use and allowing access.

---

[11] https://paris21.org/sites/default/files/Ghana NSDS 2.pdf
[12] http://opendatatoolkit.worldbank.org/en/odra.html
[13] The guide is available from https://statswiki.unece.org/display/SFSDG/Task+Force+on+National+Reporting+Platforms?preview=/128451803/170164503/NRP_practical%20guide_Note%20from%20UNCES%20SG%20SDG%20TF%20NRP.pdf . See also the background document entitled "Principles of SDG indicator reporting and dissemination platforms and guidelines for their application", which is being submitted to the consideration of the Statistical Commission at its fiftieth session.
[14] http://od4d.net/
[15] https://opendatawatch.com/reference/iodc-2018-brochure-national-reporting-for-the-sustainable-development-goals/
[16] http://www.ihsn.org/software/disclosure-control-toolbox

The first step for anonymizing microdata is to remove personally identifiable information, such as names, addresses, social security, geo-references, and id numbers. This is done by removing the information entirely and/or by adding statistical noise to the data so that the information can't be directly linked to an individual. Addresses or geospatial coordinates should be aggregated to prevent the re-identification of individual respondent while still providing sufficiently granular location information that is useful for analysis.

Though anonymization of datasets is a good practice, it is not always enough to keep a dataset private, especially in the case of datasets with many variables. High-dimensional datasets can be joined with other datasets to reidentify participants, as was done by two computer scientists for a Data for Development Challenge.[17] Extra care should be taken to anonymize and protect these high-dimensional datasets. There may remain, however, some risk of disclosure of information regardless of the steps taken. Because all methods of anonymization degrade the information contained in a dataset (and not publishing removes all value), a decision to anonymize data or limit their release must also consider the likelihood of disclosure, the harm done in case of disclosure, and the public's right to information.

Open data risk assessments, like the one that the city of Seattle implemented in 2018,[18] can be used to analyse the risks associated with different datasets and create appropriate policies to protect those data depending on the value of the data and the potential threat to their confidentiality. Open data risks assessments also help define accountability in case of breach of confidentiality. In addition, tools are being developed that will help address this challenge.

C. Activities and capabilities that support Open Data across the official statistical system

This section provides the basic elements that promote open data within the system of official statistics. Emphasis is placed on key activities and capabilities necessary to deliver open data; as well as activities to support the use of statistics among users.

Activities

Open data aligns with the United Nations Fundamental Principles of Official Statistics. Implementing the open data approach for official statistics enhances the availability of statistical information to users who monitor the economic, demographic, social and environmental situation of a country (Principle 1 of the Fundamental Principles of Official Statistics). In addition, open data activities support important official statistical norms and standards, as well as ensure confidentiality of published data. This is underpinned by a transparent legal basis (Principles 6 and 7 of the Fundamental Principles of Public Statistics).

Although open data is mainly associated with the dissemination stage, the process of making data open has an impact on many phases of the statistical production process, from users' needs specification and survey designing to survey evaluation. Employing open data standards in the statistical practice can boost efficiency in the analysis of data sets. The use of open Application Programming Interfaces (APIs) can also be beneficial in the data collection and processing phases, especially in relation to the use of administrative data sources.

[17] https://petsymposium.org/2013/papers/sharad-deanonymization.pdf

[18] https://fpf.org/wp-content/uploads/2018/01/FPF-Open-Data-Risk-Assessment-for-City-of-Seattle.pdf

Open data provides supplementary path for the official statistical system to engage with users, but it requires close collaboration with partners and customers. Current experience has shown that open data provides new access opportunities to the public; and this has resulted in measurable improvements in the form of economic growth, employment and competitiveness[19]. An open data approach can enhance the dissemination and use of official statistics.

The creation of Open Data Strategy [20] can be a useful tool to inform the society and manage statistics dissemination. For a better understanding and interpretation of statistical data, distributed also in machine-readable formats, it is necessary to develop an appropriate metadata policy. Users of the data dissemination portals and other data dissemination channels (such as online APIs) should be able to easily search and select the data they need and all the appropriate descriptive information (metadata), including complete information on the methodology used to generate the data (Principle 3 of Fundamental Principles of Official Statistics). A clear message and readability of adopted rules help to minimise the threats related to methodological misunderstandings or confusion because of disinformation. Moreover, it can be an element that reduces doubts about maintaining the confidentiality of statistical information within open data process.

In the case of official statistics, the implementation of open data often means only changes in the format of data that are already published. Considerations such as data management, version control, anonymization, data quality, and approval mechanisms that normally bear on open data are generally addressed via the national statistical system. Thus, the relatively minor task of publishing and disseminating official statistics in an open format (in addition to whatever form they are already disseminated) could produce early benefits with only modest efforts and cost, and may be possible to achieve within the NSO's existing mandate and authority

Capabilities

The necessity of using professional standards, including for the communication and delivery of statistical data to users, is embodied in the Principle 1 of the Fundamental Principles of Official Statistics. To fulfil this requirement the development of a special set of technical capabilities is vital. Additionally, the professional implementation of the above-mentioned activities requires NSOs take an innovative approach in the scope of organization and personal capabilities.

The open data process should be perceived as the continuing development of skills and competencies. In this context, NSOs should manage three groups of capabilities which are key for a successful data opening process, namely: IT, organizational capabilities, and personal capabilities. The most essential capabilities necessary to implement open data in official statistics have already been used and developed in the statistics production for years, including capabilities in research programming, quality control, data and metadata management, as well data analytics and reporting.

[19] https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf
[20] European Data Portal

In the context of open data dissemination, while considering the view of users' needs, greater emphasis must be laid on two main groups of capabilities which are needed to embed an open data approach into the broader statistical practice. They are as follow:

1. *Open data communication* includes skills related to the various channels of communication between the statistics producers and data users. It should include clear way of data presentation, storytelling with data, high quality data journalism and data visualizations which encourage users to engage and interact with the data. Additionally, a data distribution strategy should describe the way statistical data and metadata is disseminated according to the designed open data process.

2. *Data delivery skills* concern knowledge about the creation of effective data delivery mechanisms using appropriate IT tools and technical standards, such as open data formats and well-documented open API specifications. For future activities, the recognition of the linked open data (LOD)[21] concept is a key element.

The implementation of open data is an opportunity to widen the use of data assets produced national statistical systems. Data dissemination channels, IT skills development and clear communication, which are adjusted to users' needs, are factors contributing to increasing the trust in a statistical institution and are discussed in international fora[22]. Developing capabilities related to the newest technological solutions is therefore inevitable to promote integrated, effective and user-friendly products within the official statistical system.

D. Emerging issues in the Open Data Charter

As highlighted above, the United Nations Statistical Commission's Friends of the Chair Group on the Fundamental Principles of Official Statistics (FOC-FPOS) has undertaken a comparison of the 10 United Nations Fundamental Principles of Official Statistics (FPOS) and the 6 Principles of Open Data Charter (ODC). It should be noted that this is not the first time, nor the last, that the two frameworks will be considered in a unified manner. For example, many of the underlying threads / themes throughout the International Open Data Conference 2018 (IODC18) in Buenos Aires had strong linkages to the FPOS and have been echoed in 'FPOS centred' meetings.

There has been recent consultation on refreshing the International Open Data Charter (ODC) Principles, with a relaunch scheduled for early 2019. Throughout this process several issues have emerged which are relevant for National Statistical Offices (NSO) as they investigate the implementation of principles from the Open Data Charter.

Open by default

There has been strong debate throughout the consultation of the Open Data Charter Principles about the principle "open by default". Overall it is still held as a fundamental principle needed to guide behaviour towards proactive open government and the maximisation of potential value from data; however, it is felt that the expression "open by default" needs to be more clearly defined, as to its application regarding the

[21] https://en.wikipedia.org/wiki/Linked_data
[22] http://www.paris21.org/sites/default/files/2018-08/Measuring-Statistical-Capacity-Development_Web_0.pdf

FPOS and the actions of a NSO. Further work is needed to determine if "open by default" could and should be incorporated in future FPOS.

Data sovereignty

Across the world, data sovereignty is an emerging issue, whether it be in relation to the use of social media data, the storage of data in the cloud, or the custodianship of data relating to specific communities (e.g., data on indigenous population groups). Data sovereignty typically refers to the understanding that data is subject to the laws of the nation within which it is stored. Indigenous data sovereignty describes data as subject to the laws of the nation from which it is collected. Data sovereignty needs to be considered when giving effect to principles such as "open by default" (ODC) and "equity of access" (FPOS). Feedback suggests that early, open and transparent discussions can avoid long-term misunderstandings and deliver greater value from data. Opportunities do exist through community involvement and clarity of user needs.

Data governance, management and infrastructure

Sound governance and management of data is a critical to ensuring the implementation of the Open Data Charter Principles. These principles, along with open data practices of rich metadata, open standards, and open (non-proprietary) and machine-readable formats, contribute to all data, including official statistics, being more interoperable and reusable, leading to more impact and value generation.

Adequate data governance and management at both the system and enterprise levels are necessary to ensure sustainable and reliable access to data. This is essential if we want to see new products and services built on open data and for the knowledge economy to grow. An example of a government (system) led open data policy initiative is Indonesia's Satu Data[23].

Data Privacy, Security and Confidentiality

Different organisations have different requirements relating to when they must or wish to protect the, privacy, security, and confidentiality of data so that people, households, and organisations can't be identified without their knowledge. This includes where we must or wish to protect the confidentiality of data throughout its life cycle — whenever we collect, use, store, and distribute it.

The terms privacy, security, and confidentiality are often used interchangeably, but each term has a different meaning:

- **Privacy** refers to a person's right to control the availability of data about themselves.
- **Security** refers to how an organisation stores and controls access to the data it holds.
- **Confidentiality** refers to the duty to protect data from, and about, individuals and organisations; and how we ensure that data is not made available or disclosed without authorisation.

A guide has been developed to introduce data users to the concepts of privacy, security and confidentiality, and their application, with a specific focus on open data.[24] Just as the concepts of privacy, security and confidentiality are critical to the application of FPOS, they are equally important to the introduction and sustainability of open data and supporting ODC.

[23] https://govinsider.asia/inclusive-gov/the-road-to-satu-data-indonesia/

[24] https://data.govt.nz/assets/Uploads/data-confidentiality-principles-methodology-report-oct-2018.pdf

E. Conclusions

The Open Data movement presents tools and policies that can help NSOs reduce the costs of statistical processing while improving quality, build citizens' trust in national governments through transparency, and promote economic growth through sparking innovation from increased access to new data sources. Because NSOs are the main custodians of data and official statistics, they are well positioned to take a leadership role in the adoption of open data polices within countries

In addition, as the primary organization creating, organizing, and disseminating data in most countries, NSOs are critical to the success of national open data initiatives. Further, because of their demonstration effect, it is believed that NSOs' support for open data can ripple through to other coordinating agencies. The issues outlined in this paper are a good start for NSOs to begin taking the lead on open data in their own organizations and setting off a chain reaction that could enable open data at the country and international level.

While there are many issues or challenges presented, the authors consider that good progress is being made, no challenges are insurmountable and that continued collaboration between NSOs and the open data community is critical going forward. We support the recommendation that a working group on open data be created to help facilitate continued collaboration and consider this background document as a resource for further work.

Appendix A. Categories of the Open Data Watch Openness Assessment.

Open Data Watch (ODW) has operationalized the Open Definition in its Open Data Inventory (ODIN) methodology, which assesses the coverage and openness of national statistical systems. The ODW openness assessment has five elements: 1) machine readability; 2) use of non-proprietary formats; 3) availability of multiple download options; 4) availability of metadata providing sufficient context to understand the data; and 5) open licensing. ODW's assessment methodology is available for NSOs or other statistical agencies to use for self-assessment.[25]

Machine readability: When data are made available in formats that are not machine readable, such as PDF or IMG files, users cannot easily access and modify the data, which restricts the scope of the data's use. However, PDF versions of datasets within reports can be useful, as the text in conjunction with the data gives context and may help less technical users understand the data. Machine-readable file formats such as XLS, XLSX, CSV, TXT, or JSON allow users to readily process data using a computer.

Non-proprietary formats: Data that can only be accessed through costly, proprietary software may prevent some users from accessing the data at all. Open data should be available in non-proprietary formats that can be accessed with open-source software. The XLSX format is in the public domain. However, many countries still publish data in XLS files. Although XLS files can be opened with some open-source software, such as OpenOffice and LibreOffice, the format is based on BIFF (Binary Interchange File Format), which is restricted by various licenses. The PDF format is non-proprietary, but it is not machine readable.

Download options: Bulk downloads are a key component of the Open Definition, which requires data to be "provided as a whole... and downloadable via the internet." An application programming interface (API) can also be used as a bulk export option. API's, however, typically require registration and are better suited for more technical users and scenarios when the data needs to be constantly updated. For those who do not require bulk downloads, an intuitive, natural language interface that allows them to select the data of interest will be most useful.

Metadata: Information identifying the source of data, definitions of indicators, and the dates of compilation are needed by data users to evaluate the utility of the data. The experience of ODIN assessors suggests that one of the areas countries need to prioritize is standardizing both the format and location of metadata. In many cases, metadata are scattered across a website with no logical path from the published dataset. Good metadata also facilitates interoperability of data by providing the codes, descriptors, and standards to merge a dataset with other data.

Open licensing: An open license grants permission to use data freely but may impose one or more conditions. The most common condition is attribution, which requires that the original source of the data be acknowledged. The Open Definition 2.1 lists nine requirements for an open license and seven acceptable conditions or restrictions that can be included in an open license. Open licenses or terms of use posted by statistical agencies should be consistent with any underlying statistical law and should not be superseded by other statutory or common law, including copyright and defamation laws. See Box 1 Open Data Licenses.

[25] http://odin.opendatawatch.com/Downloads/otherFiles/ODIN-2017-Methodology.pdf

Appendix B: Mapping of Fundamental Principles of Official Statistics and International Open Data Charter

The purpose of this Mapping of the Fundamental Principals of Official Statistics and International Open Data Charter and other relevant principles is to show interrelationships and where principles may be able to be integrated. The mapping activity currently includes the Open Data Charter Principles (international open data charter) and Fundamental Principles of Official Statistics to highlight similarities and gaps between the 2 sets of principles.

Observations

Some elements of the International Open Data Charter are **not** included in this mapping, as the nature of their relationships with UNFPOS is not explicit. These elements are listed below:

3.24(a) (c) (d)

5.32 (b) (f)

6.37 (d) (e)

Only Principle 4 of the UNFPOS (Prevention of misuse) is not directly covered in the International Open Data Charter, and thus is also not included this mapping.

Mapping in matrix format

Note: Open Data Charter principles and Fundamental Principles of Official Statistics are included in Appendix 3 and Appendix 4. Use these attachments and their numbered references when reading the matrix to reference the relevant points.

		International Open Data Charter (✓)					
		Principle 1 Open by default P-FPOS i P-FPOS ii P-FPOS iii	**Principle 2** Timely and compre-hensive	**Principle 3** Accessible and useable	**Principle 4** Comparable and interoperable	**Principle 5** Improved governance and citizen engagement P-FPOS i P-FPOS ii P-FPOS iii	**Principle 6** For inclusive development and innovation P-FPOS i
Fundamental Principles of Official Statistics	**Principle 1** Relevance, impartiality and equal access		✓2.19 ✓2.20 ✓2.21 (b) ✓2.21 (h)	✓3.22 ✓3.23		✓ 5.28 ✓ 5.29 ✓ 5.32 (g)	✓ 6.33 ✓ 6.34 ✓ 6.35
	Principle 2 Trust in Official statistics; professional standards and ethics	✓1.15	✓2.18 ✓2.20 ✓2.21 (h) ✓2.21 (a) ✓2.21 (b) ✓2.21 (f)	✓3.22 ✓3.23 ✓3.24 (e)		✓ 5.28	
	Principle 3 Accountability and transparency	✓ 1.17 (c) ✓1.15	✓2.21 (c) ✓2.21 (e) ✓2.21 (f) ✓2.21 (d) ✓2.21 (g)	✓3.23 ✓3.24 (e)	✓4.25 ✓4.27 (b) ✓4.27 (c)	✓ 5.30 ✓5.32 (c)	✓ 6.35
	Principle 4 Prevention of misuse						
	Principle 5 Sources of official Statistics	✓1.13	✓2.21 (a) ✓2.21 (b)	✓3.24 (b)	✓4.27 (b)	✓ 5.30	
	Principle 6 Confidentiality	✓1.16 ✓ 1.17 (e) ✓ 1.17 (f)					
	Principle 7 Legislation and regulation	✓ 1.17 (a) ✓ 1.17 (b) ✓ 1.17 (c) ✓1.15				✓ 5.32 (a)	
	Principle 8 National coordination	✓1.15 ✓ 1.17 (e)			✓4.25	✓ 5.31	✓ 6.36 ✓ 6.37 (f)
	Principle 9 Use of international standards	✓1.15 ✓ 1.17 (e)			✓4.26 ✓4.27 (a) ✓4.27 (d) ✓4.27 (e)		✓ 6.37 (b) ✓ 6.37 (c) ✓ 6.37 (f)
	Principle 10 Multilateral and international cooperation	✓ 1.17 (e)			✓4.27 (d) ✓4.27 (e)		✓ 6.36 ✓ 6.37 (a) ✓ 6.37 (b) ✓ 6.37 (f)

P-FPOS i = Preamble Fundamental Principals of Official Statistics, paragraph 1 - *Bearing in mind* the critical role of high-quality official statistical information in analysis and informed policy decision-making in support of sustainable development, peace and security, as well as for mutual knowledge and trade among the States and peoples of an increasingly connected world, demanding openness and transparency.
✓1.14, ✓2.20, ✓3.22, ✓ 5.32 (d), ✓5.32 (a), ✓5.32 (e), ✓6.37 (g)

P-FPOS ii = Preamble Fundamental Principals of Official Statistics, paragraph 2 - *Bearing in mind also* that the essential trust of the public in the integrity of official statistical systems and its confidence in statistics depend to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and to respect the rights of its members and, in this context, that the professional independence and accountability of statistical agencies are crucial. ✓2.20, ✓ 5.32 (d), ✓ 5.32 (e),

P-FPOS iii = Preamble Fundamental Principals of Official Statistics, paragraph 3 - *Stressing* that, in order to be effective, the fundamental values and principles that govern statistical work have to be guaranteed by legal and institutional frameworks and respected at all political levels and by all stakeholders in national statistical systems.
✓ 1.17 (a) ✓ 1.17 (b) ✓ 1.17 (c), ✓6.37 (g)

Written description of the mapping matrix

Fundamental Principles of Official Statistics - Preamble

i. *Bearing in mind* the critical role of high-quality official statistical information in analysis and informed policy decision-making in support of sustainable development, peace and security, as well as for mutual knowledge and trade among the States and peoples of an increasingly connected world, demanding openness and transparency.

Maps to the following International Open Data Principles:

- 1.14. We recognize that free access to, and subsequent use of, government data is of significant value to society and the economy, and that government data should, therefore, be open by default.
- 2.20. We recognize that in order to be valuable to governments, citizens, and civil society and private sector organizations, data must be comprehensive, accurate, and of high quality.
- 3.22. We recognize that opening up data enables governments, citizens, and civil society and private sector organizations to make better informed decisions.
- 5.32.a Implement oversight and review processes to report regularly to the public on the progress and impact of our open data initiatives.
- 5.32.d Engage with the Freedom of Information / Access to Information / Right to Information community to align the proactive release of open data with governments' obligation to release information on request.
- 5.32.e Engage proactively with citizens and civil society and private sector representatives to determine what data they need to effectively hold governments accountable.
- 6.37.g Empower a future generation of data innovators inside and outside government by building capacity and encouraging developers, entrepreneurs, civil society and private sector organizations, academics, media representatives, government employees, and other users to unlock the value of open data.

Fundamental Principles of Official Statistics - Preamble

ii. *Bearing in mind also* that the essential trust of the public in the integrity of official statistical systems and its confidence in statistics depend to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and to respect the rights of its members and, in this context, that the professional independence and accountability of statistical agencies are crucial.

Maps to the following International Open Data Principles:

- 2.20 We recognize that in order to be valuable to governments, citizens, and civil society and private sector organizations, data must be comprehensive, accurate, and of high.

- 5.32 d Engage with the Freedom of Information / Access to Information / Right to Information community to align the proactive release of open data with governments' obligation to release information on request.
- 5.32.e Engage proactively with citizens and civil society and private sector representatives to determine what data they need to effectively hold governments accountable.

Fundamental Principles of Official Statistics - Preamble

iii. *Stressing* that, in order to be effective, the fundamental values and principles that govern statistical work have to be guaranteed by legal and institutional frameworks and respected at all political levels and by all stakeholders in national statistical systems.

Maps to the following International Open Data Principles:

- 1.17.a Develop and adopt policies and practices to ensure that all government data is made open by default, as outlined in this Charter, while recognizing that there are legitimate reasons why some data cannot be released.
- 1.17.b Provide clear justifications as to why certain data cannot be released.
- 1.17.c Establish a culture of openness, not only through legislative and policy measures, but also with the help of training and awareness programs, tools, guidelines, and communication strategies designed to make government, civil society, and private sector representatives aware of the benefits of open data.
- 6.37.g Empower a future generation of data innovators inside and outside government by building capacity and encouraging developers, entrepreneurs, civil society and private sector organizations, academics, media representatives, government employees, and other users to unlock the value of open data.

Fundamental Principle 1. Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

Maps to the following International Open Data Principles:

- 2.19 We recognize the importance of consulting with data users, including citizens, other governments, and civil society and private sector organizations to identify which data to prioritize for release and/or improvement.
- 2.20 We recognize that in order to be valuable to governments, citizens, and civil society and private sector organizations, data must be comprehensive, accurate, and of high quality.
- 2.21.b Release high-quality open data in a timely manner, without undue delay. Data will be comprehensive and accurate, and released in accordance with prioritization that is

informed by consultations with open data users, including citizens, other governments, and civil society and private sector organizations.

- 2.21.h Be transparent about our own data collection, standards, and publishing processes by documenting these processes online.
- 3.22 We recognize that opening up data enables governments, citizens, and civil society and private sector organizations to make better informed decisions.
- 3.23 We recognize that when open data is released, it should be easily discoverable and accessible, and made available without bureaucratic or administrative barriers, which can deter people from accessing the data.
- 5.28 We recognize that the release of open data strengthens the governance of and trust in our public institutions, reinforces governments' obligation to respect the rule of law, and provides a transparent and accountable foundation to improve decision-making and enhance the provision of public services.
- 5.29 We recognize that open data encourages better development, implementation, and assessment of programs and policies to meet the needs of our citizens, and enables civic participation and better informed engagement between governments and citizens.
- 5.32.g Encourage the use of open data to develop innovative, evidence-based policy solutions that benefit all members of society, as well as empower marginalized communities.
- 6.33 We recognize the importance of openness in stimulating creativity and innovation. The more governments, citizens, and civil society and private sector organizations use open data, the greater the social and economic benefits that will be generated. This is true for government, commercial, and non-commercial uses.
- 6.34 We recognize that open data can help to identify social and economic challenges, and monitor and deliver sustainable development programs. Open data can also help meet global challenges such as poverty, hunger, climate change, and inequality.
- 6.35 We recognize that open data is, by its nature, an equitable resource that empowers all people by allowing them to access data regardless of who they are or where they live. However, we also recognize the existence of a global digital divide in regard to technological tools and expertise; this divide limits the ability of socially and economically marginalized people to access and use open data.

Fundamental Principle 2. To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

Maps to the following International Open Data Principles:

- 1.15 We acknowledge the need to promote the global development and adoption of resources, standards, and policies for the creation, use, exchange, and harmonization of open data.
- 2.18 We recognize that it may require time and human and technical resources to identify data for release or publication.

- 2.20 We recognize that in order to be valuable to governments, citizens, and civil society and private sector organizations, data must be comprehensive, accurate, and of high quality.
- 2.21.a Create, maintain, and share public, comprehensive lists of data holdings to support meaningful consultations around data prioritization, publication, and release dates.
- 2.21.b Release high-quality open data in a timely manner, without undue delay. Data will be comprehensive and accurate, and released in accordance with prioritization that is informed by consultations with open data users, including citizens, other governments, and civil society and private sector organizations.
- 2.21.f Apply consistent information lifecycle management practices, and ensure historical copies of datasets are preserved, archived, and kept accessible as long as they retain value.
- 2.21.h Be transparent about our own data collection, standards, and publishing processes by documenting these processes online.
- 3.22 We recognize that opening up data enables governments, citizens, and civil society and private sector organizations to make better informed decisions.
- 3.23 We recognize that when open data is released, it should be easily discoverable and accessible, and made available without bureaucratic or administrative barriers, which can deter people from accessing the data.
- 3.24.e Ensure data can be accessed and used effectively by the widest range of users. This may require the creation of initiatives to raise awareness of open data, promote data literacy, build capacity for effective use of open data, and ensure citizen, community, and civil society and private sector representatives have the tools and resources they need to effectively understand how public resources are used.
- 5.28 We recognize that the release of open data strengthens the governance of and trust in our public institutions, reinforces governments' obligation to respect the rule of law, and provides a transparent and accountable foundation to improve decision-making and enhance the provision of public services.
-

Fundamental Principle 3. To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

Maps to the following International Open Data Principles:

- 1.15 We acknowledge the need to promote the global development and adoption of resources, standards, and policies for the creation, use, exchange, and harmonization of open data.
- 1.17.c Establish a culture of openness, not only through legislative and policy measures, but also with the help of training and awareness programs, tools, guidelines, and communication strategies designed to make government, civil society, and private sector representatives aware of the benefits of open data;
- 2.21.c To the extent possible, release data in its original, unmodified form, and link data to any relevant guidance, documentation, visualizations, or analyses.
- 2.21.d To the extent possible, release data that is disaggregated to the lowest levels of administration, including disaggregation by gender, age, income, and other categories.

- 2.21.e Allow users to provide feedback, and continue to make revisions to ensure data quality is improved as necessary;
- 2.21.f Apply consistent information lifecycle management practices, and ensure historical copies of datasets are preserved, archived, and kept accessible as long as they retain value.
- 2.21.g Consult data users on significant changes to the structure or supply of data in order to minimize the impact to users that have created tools based on open data.
- 2.21.h Be transparent about our own data collection, standards, and publishing processes by documenting these processes online.
- 3.23 We recognize that when open data is released, it should be easily discoverable and accessible, and made available without bureaucratic or administrative barriers, which can deter people from accessing the data.
- 3.24.e Ensure data can be accessed and used effectively by the widest range of users. This may require the creation of initiatives to raise awareness of open data, promote data literacy, build capacity for effective use of open data, and ensure citizen, community, and civil society and private sector representatives have the tools and resources they need to effectively understand how public resources are used.
- 4.25 We recognize that in order to be most effective and useful, data should be easy to compare within and between sectors, across geographic locations, and over time.
- 4.27.b Ensure that open datasets include consistent core metadata and are made available in human- and machine-readable formats.
- 4.27.c Ensure that data is fully described, that all documentation accompanying data understand the source, strengths, weaknesses, and analytical limitations of the data.
- 5.30 We recognize that engagement and consultation with citizens and civil society and private sector organizations can help governments understand which types of data are in high demand, and, in turn, can lead to improved data prioritization, release, and standardization practices.
- 5.32.c Provide training programs, tools, and guidelines designed to ensure government employees are capable of using open data effectively in policy development processes.
- 6.35 We recognize that open data is, by its nature, an equitable resource that empowers all people by allowing them to access data regardless of who they are or where they live. However, we also recognize the existence of a global digital divide in regard to technological tools and expertise; this divide limits the ability of socially and economically marginalized people to access and use open data.

Fundamental Principle 4. The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

(No direct mapping to Open Data Principles)

Fundamental Principle 5. Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

Maps to the following International Open Data Principles:

- 2.21.a Create, maintain, and share public, comprehensive lists of data holdings to support meaningful consultations around data prioritization, publication, and release dates.
- 2.21.b Release high-quality open data in a timely manner, without undue delay. Data will be comprehensive and accurate, and released in accordance with prioritization that is informed by consultations with open data users, including citizens, other governments, and civil society and private sector organizations.
- 3.24.b Release data in open formats to ensure that the data is available to the widest range of users to find, access, and use. In many cases, this will include providing data in multiple, standardized formats, so that it can be processed by computers and used by people.
- 4.27.b Ensure that open datasets include consistent core metadata and are made available in human- and machine-readable formats.
- 5.30 We recognize that engagement and consultation with citizens and civil society and private sector organizations can help governments understand which types of data are in high demand, and, in turn, can lead to improved data prioritization, release, and standardization practices.

Fundamental Principle 6. Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Maps to the following International Open Data Principles:

- 1.16 We recognize that open data can only be unlocked when citizens are confident that open data will not compromise their right to privacy, and that citizens have the right to influence the collection and use of their own personal data or of data generated as a result of their interactions with governments.
- 1.17.e Observe domestic laws and internationally recognized standards, in particular those pertaining to security, privacy, confidentiality, and intellectual property. Where relevant legislation or regulations do not exist or are out of date, they will be created and/or updated.
- 1.17.f In accordance with privacy legislation and standards, anonymize data prior to its publication, ensuring that sensitive, personally-identifiable data is removed.

Fundamental Principle 7. The laws, regulations and measures under which the statistical systems operate are to be made public.

Maps to International Open Data Principles

- 1.15 We acknowledge the need to promote the global development and adoption of resources, standards, and policies for the creation, use, exchange, and harmonization of open data.

- 1.17.a Develop and adopt policies and practices to ensure that all government data is made open by default, as outlined in this Charter, while recognizing that there are legitimate reasons why some data cannot be released.
- 1.17.b Provide clear justifications as to why certain data cannot be released.
- 1.17.c Establish a culture of openness, not only through legislative and policy measures, but also with the help of training and awareness programs, tools, guidelines, and communication strategies designed to make government, civil society, and private sector representatives aware of the benefits of open data.
- 5.32.a Implement oversight and review processes to report regularly to the public on the progress and impact of our open data initiatives.

Fundamental Principle 8. Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

Maps to the following International Open Data Principles:

- 1.15 We acknowledge the need to promote the global development and adoption of resources, standards, and policies for the creation, use, exchange, and harmonization of open data.
- 1.17.e Observe domestic laws and internationally recognized standards, in particular those pertaining to security, privacy, confidentiality, and intellectual property. Where relevant legislation or regulations do not exist or are out of date, they will be created and/or updated.
- 1.17.d Develop the leadership, management, oversight, performance incentives, and internal communication policies necessary to enable this transition to a culture of openness in all government departments and agencies, including official statistics organizations.
- 4.25 We recognize that in order to be most effective and useful, data should be easy to compare within and between sectors, across geographic locations, and over time.
- 5.31 We recognize that city or local governments are often the first point of interaction between citizens and government, and that these governments therefore have a crucial role in supporting citizen engagement on open data.
- 6.36 We recognize the role of governments in promoting innovation and sustainable development does not end with the release of open data. Governments must also play an active role in supporting the effective and innovative reuse of open data, and ensuring government employees, citizens, and civil society and private sector organizations have the data they need and the tools and resources to understand and use that data effectively.
- 6.37.f Build capacity and share technical expertise and experience with other governments and international organizations around the world, ensuring that everyone can reap the benefits of open data; and

Fundamental Principle 9. The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

Maps to the following International Open Data Principles:

- 1.15 We acknowledge the need to promote the global development and adoption of resources, standards, and policies for the creation, use, exchange, and harmonization of open data.
- 1.17 e Observe domestic laws and internationally recognized standards, in particular those pertaining to security, privacy, confidentiality, and intellectual property. Where relevant legislation or regulations do not exist or are out of date, they will be created and/or updated.
- 4.26 We recognize that data should be presented in structured and standardized formats to support interoperability, traceability, and effective reuse.
- 4.27.a Implement consistent, open standards related to data formats, interoperability, structure, and common identifiers when collecting and publishing data.
- 4.27 d Engage with domestic and international standards bodies and other standard setting initiatives to encourage increased interoperability between existing international standards, support the creation of common, global data standards where they do not already exist, and ensure that any new data standards we create are, to the greatest extent possible, interoperable with existing standards.
- 4.27.e Map local standards and identifiers to emerging globally agreed standards and share the results publicly.
- 6.37.b Create or explore potential partnerships between governments and with civil society and private sector organizations and multilateral institutions to support the release of open data and maximize the impact of data through effective use.
- 6.37.c Create or support programs and initiatives that foster the development or co-creation of datasets, visualizations, applications, and other tools based on open data.
- 6.37 f Build capacity and share technical expertise and experience with other governments and international organizations around the world, ensuring that everyone can reap the benefits of open data.

Fundamental Principle 10. Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

Maps to International Open Data Principles

- 1.17.e Observe domestic laws and internationally recognized standards, in particular those pertaining to security, privacy, confidentiality, and intellectual property. Where relevant legislation or regulations do not exist or are out of date, they will be created and/or updated.
- 4.27.d Engage with domestic and international standards bodies and other standard-setting initiatives to encourage increased interoperability between existing international standards, support the creation of common, global data standards where they do not already exist, and ensure that any new data standards we create are, to the greatest extent possible, interoperable with existing standards.
- 4.27.e Map local standards and identifiers to emerging globally agreed standards and share the results publicly.

- 6.36 We recognize the role of governments in promoting innovation and sustainable development does not end with the release of open data. Governments must also play an active role in supporting the effective and innovative reuse of open data, and ensuring government employees, citizens, and civil society and private sector organizations have the data they need and the tools and resources to understand and use that data effectively.
- 6.37.a Encourage citizens, civil society and private sector organizations, and multilateral institutions to open up data created and collected by them in order to move toward a richer open data ecosystem with multiple sources of open data.
- 6.37.b Create or explore potential partnerships between governments and with civil society and private sector organizations and multilateral institutions to support the release of open data and maximize the impact of data through effective use.
- 6.37.f Build capacity and share technical expertise and experience with other governments and international organizations around the world, ensuring that everyone can reap the benefits of open data.

Appendix C - International Open Data Charter

Principle 1 – Open by Default

1.13 We recognize that the term "government data" includes, but is not limited to, data held by national, regional, local, and city governments, international governmental bodies, and other types of institutions in the wider public sector. The term government data could also apply to data created for governments by external organizations, and data of significant benefit to the public that is held by external organizations and related to government programs and services (e.g. data on extractives entities, data on transportation infrastructure, etc.).

1.14 We recognize that free access to, and subsequent use of, government data is of significant value to society and the economy, and that government data should, therefore, be open by default.

1.15 We acknowledge the need to promote the global development and adoption of resources, standards, and policies for the creation, use, exchange, and harmonization of open data.

1.16 We recognize that open data can only be unlocked when citizens are confident that open data will not compromise their right to privacy, and that citizens have the right to influence the collection and use of their own personal data or of data generated as a result of their interactions with governments.

1.17 We will

a. Develop and adopt policies and practices to ensure that all government data is made open by default, as outlined in this Charter, while recognizing that there are legitimate reasons why some data cannot be released;
b. Provide clear justifications as to why certain data cannot be released;
c. Establish a culture of openness, not only through legislative and policy measures, but also with the help of training and awareness programs, tools, guidelines, and communication strategies designed to make government, civil society, and private sector representatives aware of the benefits of open data;
d. Develop the leadership, management, oversight, performance incentives, and internal communication policies necessary to enable this transition to a culture of openness in all government departments and agencies, including official statistics organizations;
e. Observe domestic laws and internationally recognized standards, in particular those pertaining to security, privacy, confidentiality, and intellectual property. Where relevant legislation or regulations do not exist or are out of date, they will be created and/or updated; and
f. In accordance with privacy legislation and standards, anonymize data prior to its publication, ensuring that sensitive, personally-identifiable data is removed.

Principle 2 – Timely and comprehensive

2 18 We recognize that it may require time and human and technical resources to identify data for release or publication.

2.19 We recognize the importance of consulting with data users, including citizens, other governments, and civil society and private sector organizations to identify which data to prioritize for release and/or improvement.

2.20 We recognize that in order to be valuable to governments, citizens, and civil society and private sector organizations, data must be comprehensive, accurate, and of high quality.

2.21 We will

a. Create, maintain, and share public, comprehensive lists of data holdings to support meaningful consultations around data prioritization, publication, and release dates;
b. Release high-quality open data in a timely manner, without undue delay. Data will be comprehensive and accurate, and released in accordance with prioritization that is informed by consultations with open data users, including citizens, other governments, and civil society and private sector organizations;
c. To the extent possible, release data in its original, unmodified form, and link data to any relevant guidance, documentation, visualizations, or analyses;
d. To the extent possible, release data that is disaggregated to the lowest levels of administration, including disaggregation by gender, age, income, and other categories;
e. Allow users to provide feedback, and continue to make revisions to ensure data quality is improved as necessary;
f. Apply consistent information lifecycle management practices, and ensure historical copies of datasets are preserved, archived, and kept accessible as long as they retain value;
g. Consult data users on significant changes to the structure or supply of data in order to minimize the impact to users that have created tools based on open data; and
h. Be transparent about our own data collection, standards, and publishing processes by documenting these processes online.

Principle 3 – Accessible and Usable

3.22 We recognize that opening up data enables governments, citizens, and civil society and private sector organizations to make better informed decisions.

3.23 We recognize that when open data is released, it should be easily discoverable and accessible, and made available without bureaucratic or administrative barriers, which can deter people from accessing the data.

3.24 We will:

a. Publish data on a central portal, so that open data is easily discoverable and accessible in one place;
b. Release data in open formats to ensure that the data is available to the widest range of users to find, access, and use. In many cases, this will include providing data in multiple, standardized formats, so that it can be processed by computers and used by people;
c. Release data free of charge, under an open and unrestrictive licence;
d. Release data without mandatory registration, allowing users to choose to download data without being required to identify themselves; and

e. Ensure data can be accessed and used effectively by the widest range of users. This may require the creation of initiatives to raise awareness of open data, promote data literacy, build capacity for effective use of open data, and ensure citizen, community, and civil society and private sector representatives have the tools and resources they need to effectively understand how public resources are used.

Principle 4 – Comparable and interoperable

4.25 We recognize that in order to be most effective and useful, data should be easy to compare within and between sectors, across geographic locations, and over time.

4.26 We recognize that data should be presented in structured and standardized formats to support interoperability, traceability, and effective reuse.

4.27 We will:

a. Implement consistent, open standards related to data formats, interoperability, structure, and common identifiers when collecting and publishing data;
b. Ensure that open datasets include consistent core metadata and are made available in human- and machine-readable formats;
c. Ensure that data is fully described, that all documentation accompanying data understand the source, strengths, weaknesses, and analytical limitations of the data;
d. Engage with domestic and international standards bodies and other standard setting initiatives to encourage increased interoperability between existing international standards, support the creation of common, global data standards where they do not already exist, and ensure that any new data standards we create are, to the greatest extent possible, interoperable with existing standards; and
e. Map local standards and identifiers to emerging globally agreed standards and share the results publicly.

Principle 5 – For improved governance and citizen engagement

5.28 We recognize that the release of open data strengthens the governance of and trust in our public institutions, reinforces governments' obligation to respect the rule of law, and provides a transparent and accountable foundation to improve decision-making and enhance the provision of public services.

5.29 We recognize that open data encourages better development, implementation, and assessment of programs and policies to meet the needs of our citizens and enables civic participation and better informed engagement between governments and citizens.

5.30 We recognize that engagement and consultation with citizens and civil society and private sector organizations can help governments understand which types of data are in high demand, and, in turn, can lead to improved data prioritization, release, and standardization practices.

5.31 We recognize that city or local governments are often the first point of interaction between citizens and government, and that these governments therefore have a crucial role in supporting citizen engagement on open data.

5.32 We will:

a. Implement oversight and review processes to report regularly to the public on the progress and impact of our open data initiatives;
b. Ensure that information published as a result of transparency or anticorruption laws is released as open data;
c. Provide training programs, tools, and guidelines designed to ensure government employees are capable of using open data effectively in policy development processes;
d. Engage with the Freedom of Information / Access to Information / Right to Information community to align the proactive release of open data with governments' obligation to release information on request;
e. Engage proactively with citizens and civil society and private sector representatives to determine what data they need to effectively hold governments accountable;
f. Respect citizens' right to freedom of expression by protecting those who use open data to identify corruption or criticize governments; and
g. Encourage the use of open data to develop innovative, evidence-based policy solutions that benefit all members of society, as well as empower marginalized communities.

Principle 6 – For inclusive development and innovation

6.33 We recognize the importance of openness in stimulating creativity and innovation. The more governments, citizens, and civil society and private sector organizations use open data, the greater the social and economic benefits that will be generated. This is true for government, commercial, and non-commercial uses.

6.34 We recognize that open data can help to identify social and economic challenges, and monitor and deliver sustainable development programs. Open data can also help meet global challenges such as poverty, hunger, climate change, and inequality.

6.35 We recognize that open data is, by its nature, an equitable resource that empowers all people by allowing them to access data regardless of who they are or where they live. However, we also recognize the existence of a global digital divide in regard to technological tools and expertise; this divide limits the ability of socially and economically marginalized people to access and use open data.

6.36 We recognize the role of governments in promoting innovation and sustainable development does not end with the release of open data. Governments must also play an active role in supporting the effective and innovative reuse of open data, and ensuring government employees, citizens, and civil society and private sector organizations have the data they need and the tools and resources to understand and use that data effectively.

6.37 We will:

a. Encourage citizens, civil society and private sector organizations, and multilateral institutions to open up data created and collected by them in order to move toward a richer open data ecosystem with multiple sources of open data;
b. Create or explore potential partnerships between governments and with civil society and private sector organizations and multilateral institutions to support the release of open data and maximize the impact of data through effective use;
c. Create or support programs and initiatives that foster the development or co-creation of datasets, visualizations, applications, and other tools based on open data;
d. Engage with schools and post-secondary education institutions to support increased open data research and to incorporate data literacy into educational curricula;
e. Conduct or support research on the social and economic impacts of open data;
f. Build capacity and share technical expertise and experience with other governments and international organizations around the world, ensuring that everyone can reap the benefits of open data; and
g. Empower a future generation of data innovators inside and outside government by building capacity and encouraging developers, entrepreneurs, civil society and private sector organizations, academics, media representatives, government employees, and other users to unlock the value of open data.

Appendix D - Fundamental Principles of Official Statistics

Principle 1. Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

Principle 2. To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

Principle 3. To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

Principle 4. The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

Principle 5. Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

Principle 6. Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Principle 7. The laws, regulations and measures under which the statistical systems operate are to be made public.

Principle 8. Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

Principle 9. The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

Principle 10. Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

ANNEX

DATA INTEROPERABILITY: A PRACTITIONER'S GUIDE TO JOINING UP DATA IN THE DEVELOPMENT SECTOR

Table of Contents

Table of Figures

Acknowledgements

This Guide was authored by Luis Gerardo González Morales, Statistician at the United Nations Statistics Division (UNSD); and, Tom Orrell, Founder of DataReady Limited on behalf of the Global Partnership for Sustainable Development Data (GPSDD).

The Guide was supported by all members of the Collaborative on SDG Data Interoperability. On behalf of UNSD and GPSDD, the authors would like to thank in particular: Shaida Badiee, Eric Swanson and their team at Open Data Watch; Enrique Jesus Ordaz López and his team at the National Institute of Statistics and Geography (INEGI) in Mexico; Bill Anderson and Beata Lisowska at Development Initiatives; Joshua Powell at Development Gateway; Radhika Lal at the United Nations Development Programme (UNDP); Jonathan Challener and his colleagues at the Organization for Economic Co-operation and Development (OECD), Malarvizhi Veerappan and her colleagues at the World Bank; and, Rose Aiko from the GPSDD Secretariat for their support and inputs into the many drafts that were produced.

We would also like to especially thank Larry Sperling, Mark DeZalia and their colleagues in PEPFAR at the U.S. State Department for their contributions and insights; as well as Claire Melamed, CEO at GPSDD and Francesca Perucci, Assistant Director at UNSD for their oversight and continued commitment to the Collaborative on SDG Data Interoperability.

Acronyms

ADX	Aggregate Data Exchange
ADX-HIV	Aggregate Data Exchange for HIV
API	Application Programming Interface
ART	Antiretroviral Therapy
CKAN	Comprehensive Knowledge Archive Network
CMF	Common Metadata Framework
CODATA	Committee on Data of the International Council for Science
CSV	Comma-Separated Values
DAMA DM-BK	Data Management Body of Knowledge
DCAT	Data Catalogue (data vocabulary)
DGO	Data Governance Officer
DSD	Data Structure Definition (in SDMX)
DST	Data Stewardship Team
EDIFACT	Electronic Data Interchange for Administration, Commerce and Transport
EO	Earth Observations
ETL	Extract-Transform-Load
GPSDD	Global Partnership for Sustainable Development Data
GSBPM	Generic Statistical Business Process Model
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immune Deficiency Syndrome
HL7	Health Level Seven
HS	Harmonized System
HTTP	HyperText Transfer Protocol
ICD	International Classification of Diseases
ICT	Information Communications Technology
IGOs	Inter-Governmental Organizations
ISO	International Organization for Standardization
JSON	JavaScript Object Notation

MDAs	Ministries, Departments and Agencies
MOU	Memorandum of Understanding
M&E	Monitoring and Evaluation
NGOs	Non-Governmental Organizations
NSOs	National Statistical Offices
OAS	OpenAPI Specification
OECD	Organization for Economic Co-operation and Development
OGC	Open Geospatial Consortium
OLAP	Online Analytical Processing
OWL	Web Ontology Language
RDF	Resource Description Framework
SDGs	Sustainable Development Goals
SDMX	Standard Data and Metadata Exchange
SKOS	Simple Knowledge Organization System
SNOMED	Systematized Nomenclature of Medicine
UML	Unified Modelling Language
UNECE	United Nations Economic Commission for Europe
UNSC	United Nations Statistics Commission
UNSD	United Nations Statistics Division
UNWDF	United Nations World Data Forum
URI	Uniform Resource Identifier
WHO	World Health Organization
W3C	World Wide Web Consortium
XML	Extensible Mark-up Language

Foreword

Background

The first United Nations World Data Forum (UNWDF) took place in Cape Town, South Africa, in January 2017. At that Forum, the foundations were laid for a new joint endeavour to explore opportunities and identify good practices for enhancing data interoperability in the area of sustainable development. These aims are today embodied in the Collaborative on SDG Data Interoperability. The Collaborative was formally established at a side-event to the 48th UN Statistical Commission (UNSC) which took place in March 2017. It is convened by the UN Statistical Division (UNSD) and Global Partnership for Sustainable Development Data (GPSDD).

Over the past two years, the Collaborative has grown significantly in size, with over 90 individuals representing entities from across the data for development spectrum – from official statistics representatives to local civil society groups – now engaged in its processes, discussions and events. The Collaborative is founded on the belief that data interoperability can: a) help to generate better quality and more holistic information that can facilitate in the achievement and monitoring of progress toward the Sustainable Development Goals (SDGs); and b) help to foster more coordinated, coherent and data-driven cooperation and collaboration between official statistical entities and the broader data ecosystem.

At the Data for Development Festival in Bristol in March 2017, the Collaborative agreed to produce guidance on data interoperability for development practitioners. This document is the first attempt at producing such guidance. It is our hope that it provides a useful starting point for statisticians, government officials, development practitioners responsible for data management, as well as suppliers of information and communication technologies (ICT) solutions in the development sector. As this Guide will further explain, interoperability is both a characteristic of good quality data and a concept that can be used to help frame data management policies.

How to use this Guide

The Guide is structured around five areas that the Collaborative has collectively identified as being integral to the development of more interoperable data systems at scale over time:

1. Interoperability, data management, and governance;
2. Canonical data and metadata models;
3. Classifications and vocabularies;
4. Standardized interfaces; and
5. Linked data.

The five areas covered by the Guide address some of the key dimensions needed to scale interoperability solutions to macroscopic and systemic levels. The Guide has been developed as a practical tool to help improve the integration and reusability of data and data systems. New sections, examples and guidance will be added to the Guide over time to ensure its continued relevance and usefulness in this fast-evolving space. Not all chapters will be relevant to all audience groups. We envisage that the introduction and first chapter will be most relevant to those engaged in policy, management and planning work; with the remaining four chapters being most relevant to technical specialists and statisticians across stakeholder

groups who are looking for specific guidance on how to improve the interoperability of their information systems.

The Guide aims for clarity and accessibility while simultaneously exploring technically complex issues. This is a difficult balance to strike but one that we have striven to maintain throughout, in some areas probably more successfully than others. It is our hope that this corpus of knowledge and examples will grow in time as the Guide matures from this first edition.

Each chapter concludes with sections entitled 'Building a Roadmap' and 'Further Reading'. These are key components of the Guide's practical application. Collectively, the Roadmap components set out an assessment framework that data managers in development organizations and government Ministries Departments and Agencies (MDAs) can use to assess the degree to which their systems are interoperable or not and where further action is required (see Annex A for further information). As with the Guide in general, it is hoped that this assessment tool will be developed further in the coming years and applied by organizations and institutions across stakeholder groups (drawing lessons from, and building on, sectoral initiatives such as the Health Data Collaborative's Health Information Systems Interoperability Maturity Toolkit[1]).

The Collaborative on SDG Data Interoperability will continue to build and maintain the Guide as it develops as a tool. Focus will shift to the development of additional modules and examples for the Guide as well as the production of ancillary materials to help raise awareness of its existence and usability. 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 a consistent approach to the system-wide improvement of data interoperability can start to become a reality in the development sector.

To find out more about the Collaborative on SDG Data Interoperability and how to contribute to the next iteration of this Guide, please contact info@data4sdgs.org.

[1]For further information see: https://www.measureevaluation.org/resources/tools/health-information-systems-interoperability-toolkit

Introduction

Over the years, countless systems that do not talk to one another have been created within and across organizations for the purposes of collecting, processing and disseminating data for development. With the proliferation of different technology platforms, data definitions and institutional arrangements for managing, sharing and using data, it has become increasingly necessary to dedicate resources to integrate the data necessary to support policy-design and decision-making.

Interoperability is the ability to join-up and merge data without losing meaning (JUDS 2016). In practice, data is said to be interoperable when it can be easily re-used and processed in different applications, allowing different information systems to work together. Interoperability is a key enabler for the development sector to become more data-driven.

In today's world, people's expectations are for greater interconnectivity and seamless interoperability, so different systems can deliver data to those who need it, in the form they need it. Data interoperability and integration[2] are therefore crucial to data management strategies in every organization. However, teams and organizations are often overloaded with day-to-day operations, and have little time left to introduce and adopt standards, technologies, tools and practices for greater data interoperability. Within the process of devising such strategies, exploring and adopting conceptual frameworks can help practitioners to better organize ideas and set the scene for the development of more tailored, detailed, and interoperable approaches to data management.

Interoperability as a conceptual framework

Interoperability is a *characteristic* of good quality data, and it relates to broader concepts of value, knowledge creation, collaboration, and fitness-for-purpose. As one of the interviewees in *The Frontiers of Data Interoperability for Sustainable Development* put it, "the problem with interoperability is that it... means different things to different people." (JUDS 2016, 5). Part of the reason for this is that interoperability exists in varying degrees and forms, and interoperability issues need to be broken down into their key components, so that they can be addressed with concrete, targeted actions.

Conceptual frameworks help us to consider interoperability in different contexts and from different perspectives. For instance:

- from a diversity of technological, semantic, or institutional viewpoints, recognizing that interoperability challenges are multi-faceted and manifest in different ways across scenarios and use cases; and
- within the context of the data value chain, as well as within the context of broader data ecosystems.

[2] While the focus of this Guide is on data interoperability, it is also important to highlight its close connection to data 'integration' which is the act of incorporating two or more datasets into the same system in a consistent way. Data integration is one of the possible outcomes of data interoperability.

FIGURE 1: DATA COMMONS FRAMEWORK

Following the *Data Commons Framework* devised by Goldstein et al (2018), we can split out the concept of interoperability into a number of narrow and broad layers that relate to standardization and semantics respectively. These layers can help in the development of projects, plans, and roadmaps to better understand interoperability needs at various points and can be summarised thus:

1. **Technology layer**: This represents the most basic level of data interoperability, and is exemplified by the requirement that data be published, and made accessible through standardized interfaces on the web;
2. **Data and format layers:** These capture the need to structure data and metadata according to agreed models and schemas, and to codify data using standard classifications and vocabularies;
3. **Human layer**: This refers to the need for a common understanding among users and producers of data regarding the meaning of the terms used to describe its contents and its proper use (there is an overlap here with the technology and data layers, in that the development and use of common classifications, taxonomies, and ontologies to understand the semantic relationships between different data elements are crucial to machine-to-machine data interoperability);
4. **Institutional and organisational layers:** These are about the effective allocation of responsibility (and accountability) for data collection, processing, analysis and dissemination both within and across organizations. They cover aspects such as data sharing agreements, licenses, and memoranda of understanding (see Annex B for more detail on legal frameworks).

These various 'layers' of interoperability are explored throughout the Guide and manifest in various ways. They also provide a useful frame of reference when thinking about interoperability needs at a systemic scale; as the example in Figure 2 demonstrates.

FIGURE 2: A USER-CENTRIC APPROACH TO INTEROPERABILITY AND OPEN DATA

Many National Statistical Offices (NSOs) are now adopting open data policies that authorize and facilitate the reuse of their statistical products, including sometimes the datasets relied upon to produce them. When thinking about how to openly publish data, it is crucial to identify the different needs of the various audiences that are likely to want to use that information.

For instance, analysts may want to access multiple datasets in machine-readable format, so they can be easily fed into statistical models to test hypotheses or make predictions. Similarly, application developers may want to use Application Programming Interfaces (APIs) that provide online access to data in standardized, open formats, so they can build interactive dashboards, maps and visualizations.

In contrast, journalists and policy makers are more likely to want to access the data in human readable formats such as tables, charts and maps, and to appreciate the ability to explore and discover related data and information using web search engines.

Each of these prospective use cases requires data interoperability at various junctures.

For the developers, having the data published electronically in digital formats is the most basic and fundamental interoperability requirement. Thereafter, having the data published in open machine-readable data formats such as the eXtensible Mark-up Language (XML), Comma Separated Values (CSV) format or JavaScript Object Notation (JSON) is a crucial next step. In more advanced systems, having metadata available using common vocabularies and adhering to commonly used metadata schemes (e.g., the Data Catalogue (DCAT) schema), is a bonus.

Journalists and researchers, on the other hand, are more interested in the ability to analyze, group and compare various datasets along meaningful categories. In other words, they are interested in the semantic coherence and comparability of data. This requires ensuring that the published data conforms to standard methods, definitions and classifications across countries and institutions.

Underpinning these use cases is a need for clear and agreed rules for accessing, using and re-using data from different sources. In this context, 'reuse' requires data to be interoperable not only from a technical perspective, but also from a legal and institutional perspective (including so-called 'legal interoperability', which forms the basis for cross-jurisdictional data sharing and use).

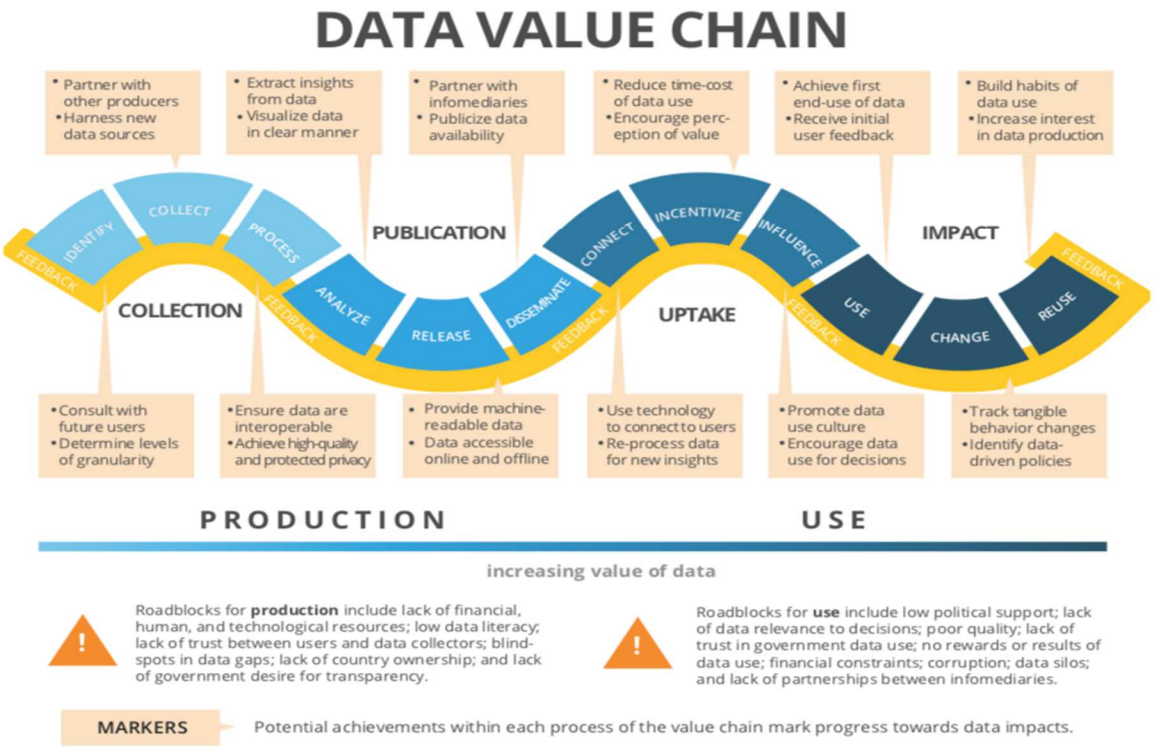
Another point that it is important to keep in mind when thinking about interoperability is that maximum levels of interoperability are not always desirable and can in fact be harmful or even unlawful (e.g., if they result in the unintentional disclosure of personal data). Before any decisions can be made on the degree to which a dataset should be made interoperable, careful consideration should be given to what the intended and anticipated use case of a dataset or IT system will be.

> "One of the primary benefits of interoperability is that it can preserve key elements of diversity while ensuring that systems work together in ways that matter most. One of the tricks to the creation of interoperable systems is to determine what the optimal level of interoperability is: in what ways should the systems work together, and in what ways should they not?" (Palfrey et al 2012, p 11).

Interoperability across the data lifecycle and value chain

A useful framework for development practitioners seeking to better understand how interoperability can add value to data is the idea of the *Data Value Chain* (Open Data Watch 2018), which highlights the role that interoperability plays in binding together its various components.

FIGURE 3: THE DATA VALUE CHAIN



Within this model, interoperability is explicitly referenced as part of the processing stage of data collection; for example, ensuring that the right classifications and standards are used to collect and record data from the outset or that the individuals tasked with collecting data have liaised with counterparts in other organizations to define how they will capture and store it. The message here is two-fold: on the one hand, planning for interoperability during the data collection stage of a dataset's lifecycle is an important part of thinking about prospective use cases down the line. At the same time, how datasets are used should also inform what steps are taken at the data collection stage so that needs are anticipated, and processes optimized. Interoperability, therefore, should be linked both to data collection and use within programmatic cycles, and this should be reflected in organizational practices and data management plans that cover the full breadth of the value chain.

Data interoperability and the SDGs

In 2015, all UN member states adopted 17 Sustainable Development Goals (SDGs) as part of the 2030 Agenda for Sustainable Development (the 2030 Agenda), which spans socio-economic development, environmental protection, and tackling economic inequalities on a global scale. The unprecedented scope

and ambition of the 2030 Agenda requires the design, implementation and monitoring of evidence-based policies using the best data and information available from multiple sources – including administrative data across the national statistical system, and data ecosystems more broadly. In this context, the Data Commons Framework introduced in the previous section can help us to understand the nature of the many data interoperability challenges that need to be addressed to support evidence-based decision making to achieve the SDGs.

Because the sustainable development field is global – the 'indivisible', 'holistic', and 'universal' dimensions of the 2030 Agenda are some of its core attributes – it is not enough for development professionals to have a common understanding of the language of sustainable development. Government Ministries Departments and Agencies (MDAs), National Statistics Offices (NSOs), intergovernmental organizations (IGOs) including UN agencies, non-governmental organizations (NGOs), and other interest groups all need to interpret and share data and information in a way that is logical and makes sense.

Sharing data and information in the sustainable development field necessitates having a common understanding of the semantics used by all groups of stakeholders involved. For example, to understand the impact of climate change on macro-economic trends, development economists must learn and understand the meaning of a host of scientific terms to understand how a changing climate will impact economic indicators. 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 combined with data gleaned from satellite and other observational and sensory sources and sometimes further reinforced by perceptions data generated by citizens themselves (so-called citizen-generated data, or CGD). Common ways of organizing data, and information are needed to enable the exchange of knowledge between policy makers and development practitioners.

Another component to realizing effective data sharing, and particularly common semantics, is the use of industry standards. Across a number of sectors, there are both information models and accepted terminologies/coding systems, which provide the semantic foundation for the sharing of information. Key to this sharing is the ability to not only share labels, but to maintain consistency of meaning, particularly across organizations or national boundaries. For example, within the healthcare domain, terminologies such as the Systematized Nomenclature of Medicine (SNOMED) provide millions of well-defined concepts and their interrelationships, reducing ambiguity in the practice of clinical medicine and documentation of patient observations in their medical records (U.S. National Library of Medicine 2018).

From a data perspective, the SDG data ecosystem is characterized by several tensions:

- between global and local data needs – for instance between globally comparable statistics and disaggregated data that is compiled for local decision-making;
- between top-down data producers (such as UN agencies or multilateral and bilateral development entities) and bottom-up ones such as small civil society organizations or local companies;
- between structured data exchange processes, such those based on the Statistical Data and Metadata eXchange (SDMX) suite of standards, and more organic processes, such as informal in-country data sharing between development actors; and
- between data producers and users from sectoral (health, education, etc.) and cross-cutting (gender, human-rights, partnerships) domains.

Within this complex matrix of processes and different levels of capacity and resources available for investment in data, coordination is key. In situations where coordination is weak, information systems and data platforms often do not share common goals and miss opportunities to create synergies and coherence. For example, even within individual NSOs or government MDAs, different IT solution providers contracted separately as part of different programmes of work or donor-sponsored projects may end up creating siloed information systems that produce architectures and datasets that are not interoperable and whose data outputs cannot be integrated with each other. This is a common challenge that directly inhibits the efficient production and processing of data needed to achieve and monitor the SDGs. Resolving the problem requires a coordinated approach and set of common guidelines across governments that consider interoperability from the outset when it comes to the procurement of IT solutions. This requires ensuring that data management and governance principles become integral components of organizational strategies and business processes. At a more systemic level, it may also mean taking a leaf out of the book of international standard development organizations such as the World Wide Web Consortium (W3C), the International Organization for Standardization (ISO), the Open Geospatial Consortium (OGC), and others.

In sum, interoperability is an *approach* that can help the development sector leverage the potential of data to increase the socio-economic value of the outcomes it is working towards. A data governance framework is needed to ensure interoperability exists between the data we collectively need to collect to both inform policies to achieve the SDGs and measure our progress is doing so.

Chapter 1: Data Management, Governance and Interoperability

"Successful data management must be business-driven, rather than IT driven." (DAMA 2017)

Overview

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 effectively governed, controlled with oversight and accountability, as it moves within and between organizations during its lifecycle.

As it stands, when entities relegate anything that has to do with 'data' to their IT or Monitoring and Evaluation (M&E) departments, without also focusing on data issues at a leadership level, they miss an opportunity. This is because they are failing to make the connection between 'data' and the sources of information that programmatic specialists – public health experts, education specialists, protection officers, natural scientists, etc. – rely on to perform their jobs, devising and implementing development policies, programmes and projects to help meet the targets of the 2030 Agenda for Sustainable Development.

Data issues need to be considered as cross-cutting, in the same way that gender, human rights and partnerships' issues currently are in the development field. As such, they require far more cogent management, funding, oversight and coordination than they are currently afforded.

This section explores the concepts of data interoperability and integration, management and governance in more detail; highlighting some useful institutional tools and examples that can help practitioners in the development of their data management and governance strategies. It sets out the various institutional frameworks and models of data governance that exist, explains the need for oversight and accountability across the data value chain, and the need for effective legal and regulatory frameworks.

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.

Institutional frameworks and interoperability

Institutional frameworks refer to the overarching systems of laws, strategies, policies, conventions, and business processes that shape how individuals, organizations, and institutions behave and engage with each other. Keeping the Data Commons Framework referred to in the introduction in mind, it is clear that institutional frameworks have a key role to play in creating the environment where data, technology, and business processes fit with each other and enable the effective functioning of knowledge-driven organizations.

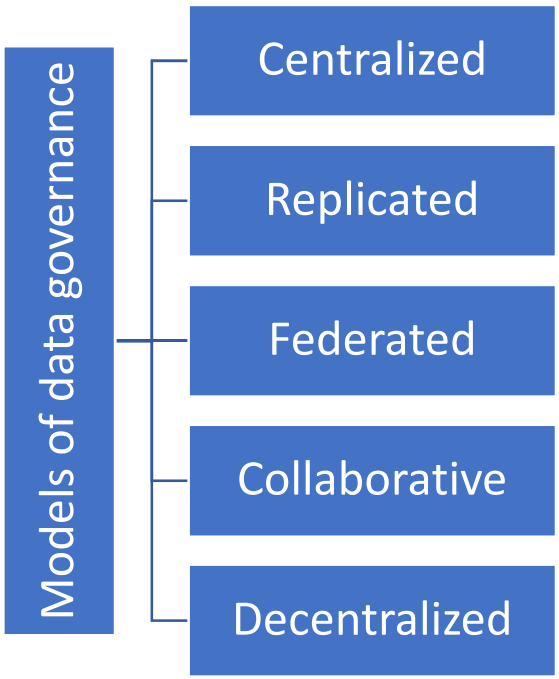
For the purposes of this Guide, we have broken down 'institutional frameworks' into three components that capture various dimensions of interoperability:

1. Institutional models of data governance;
2. Oversight and accountability models; and
3. Legal and regulatory frameworks.

Institutional models of data governance

There are different approaches to data governance, with some being more centralized than others. Individual organizations need to determine what will work best for them, keeping in mind the purpose for which data is being collected and used. For example, an NSO may want to develop a more centralized model for data collection, standard-setting, validation and security, given its role in coordinating the overall production and dissemination of official statistics at the national level. A more decentralized or more modular model of data governance may work better in instances where control over data is distributed.

FIGURE 4: MODELS OF DATA GOVERNANCE



Too much decentralization does not work well in volatile environments that require data standards and coordination to tackle global information sharing challenges. Conversely, too much centralization can hinder experimentation and the creativity needed to innovate and to respond to emerging needs of data users and the quickly changing technological landscape.

A middle ground can be found in so called "replicated" and "federated" governance frameworks. The former is when a common data governance model is adopted (usually with only minor variations) by different organizations. The latter is when multiple organizations coordinate to maintain consistency across their data governance policies, standards and procedures, although with different schedules based on their level of engagement, maturity and resources.

A replicated data governance framework is well suited to promote interoperability across independent organizations and loosely coupled data communities, each of which has ownership over specific data assets. However, this kind of governance framework requires very clear institutional and technical mechanisms for communication and collaboration, including the provision of adequate incentives for the adoption of open standards and common data and metadata models, classifications, patterns for the design of user interfaces.

A federated governance framework allows multiple departments or organizations, none of which individually controls the all the data and technological infrastructure, to constitute a decentralized but coordinated network of interconnected "hubs". Such "hubs" consolidate and provide a consistent view

FIGURE 5: FEDERATED INFORMATION SYSTEMS FOR THE SDGS

At its 49th session in March 2018, the UN UNSC welcomed the establishment of a federated system of national and global data hubs for the SDGs. By leveraging the opportunities of web technologies, this initiative is already facilitating the integration of statistical and geospatial information, promoting standards-based data interoperability and fostering collaboration among partners from different stakeholder groups to improve data flows and global reporting of the SDGs under the principles of national ownership and leadership.

The Federated Information System for the SDGs initiative has already launched various data hubs, including the global Open SDG Data Hub available from: http://www.sdg.org/.

of all the data assets available across the network, reducing the complexity of data exchange management, and provides a space where disparate members of that network can engage with one another. Moreover, although the federated model provides a coordinated framework for data sharing and communication, it also allows for multiple representations of information based on the different needs and priorities of participating data communities. It leverages technology to enable collaboration and the implementation of common data governance mechanisms.

Collaborative approaches to data governance exist between organizations and institutions and can be an effective way to engender a more multi-stakeholder, open and ecosystem approach to the tackling of interoperability problems. The benefits of collaborative approaches also include greater adaptability and flexibility than the more formal models mentioned above. The Collaborative on SDG Data Interoperability is one such example, as are the Health Data Collaborative[3], the Committee on Data of the International Council for Science (CODATA) [4] and even more formalized international standards organizations such as W3C[5] and ISO[6].

A level below that of governance frameworks sit business processes; series of tasks and activities that collectively result in the delivery of a product or service. Here, interoperability can play a role in helping gel the various parts of the business process together. The Generic Statistical Business Model (GSBPM) is a case in point.

[3] For more information see: https://www.healthdatacollaborative.org
[4] For more information see: http://www.codata.org
[5] For more information see: https://www.w3.org
[6] For more information see: https://www.iso.org/home.html

FIGURE 6: THE GENERIC STATISTICAL BUSINESS PROCESS MODEL

The GSBPM developed by the UN Economic Commission for Europe (UNECE) on behalf of the international statistical community, is an excellent example of a systemic, coordinated and collaborative initiative that has established a common standards-based approach to business process development for official statistics. The model offers examples of good practice for handling data interoperability and integration issues from a data management perspective.

As a business model, its objective is to set out and break down into logical components the tasks and activities that should take place within a statistical office to achieve organizational objectives. It, "provides a standard framework and harmonized terminology to help statistical organizations to modernize their statistical production processes, as well as to share methods and components. The GSBPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement." (Lalor 2018).

GSBPM describes itself as a reference model and makes clear that NSOs can adopt as much or as little of it as they need to and adapt its components to their own needs. It covers components across the whole statistical production chain at two levels. Within GSBPM, as in ODW's data value chain framework referenced above, interoperability and integration issues emerge explicitly at the processing stage of the model; however, are also more subtly present along the whole chain.

Taking a broad view, and keeping in mind the Data Commons Framework referred to in the introduction, dimensions of the GSBPM that are conducive to interoperability include: the specification of a Creative Commons Attribution License for reuse (see Annex B for more detail on licenses); it's standards-based nature that promotes a harmonized approach; the fact that the model considers interlinkage to other business processes from the outset; it's consideration and incorporation of a statistics-specific Common Metadata Framework (CMF); and the modular and reference nature of its components that make it possible for NSOs to align some of their functions to the common standard while retaining overall independence and the ability to choose practices that work best for them.

Oversight and accountability models

As repeated often throughout this Guide, interoperability is a characteristic of high-quality data that should be fostered across organizations; not just by computer scientists, technical experts, or IT departments within organizations. To embed data interoperability as a guiding principle across an organization requires careful planning of governance mechanisms, including appreciating the value and usefulness of oversight and accountability. The form that oversight and accountability will take depends on the size of the organization, the availability of resources, management structure, and the role of the organization in the broader data ecosystem.

Individual data governance officers (DGOs) and data stewardship teams (DSTs) (DAMA 2017, 91) should be clearly identified within operational departments, with the responsibility and commensurate authority to ensure that data is properly governed across its life cycle – that is, retains its value as an asset by being comprehensive, timely, supported by metadata, in conformity with appropriate standards, released in multiple formats for different audiences and in compliance with any applicable laws and regulations. DGO's and DST's should also bear responsibility for maintaining relationships with other organizations and entities, with a mandate to coordinate the best approaches to sharing data, keeping in mind the types of

FIGURE 7: CONFLICTING VIEWS OF DATA GOVERNANCE

Working out how to govern data within and across organizations is difficult, hence the variety of models that exists. Although this Guide suggests the approach as set out in the DAMA *Body of Knowledge* (2017), other approaches exist.

For instance, in *Non-invasive Data Governance* (Steiner 2014), Robert Steiner advocates an approach in which employees of organizations do not need to explicitly acknowledge data governance, because it is already happening; what is needed is that the process be formalized to some degree.

While approaches may differ, what is important is that organizations take a proactive approach to working out what would work best for them, in the context they work in, to effectively govern the data that flows through their systems.

conceptual, institutional and technical frameworks that will be needed to ensure interoperability across entities.

In larger organizations, a data governance council or data governance steering committee may be an appropriate mechanism to collectively govern data across its lifecycle. Any council or committee should include a mix of technical, operational and support staff and have executive support and oversight to ensure accountability. Their functions should mirror the same functions as DGOs and DSTs. This format reflects several dimensions of interoperability: technical and data, semantic, and institutional and will ensure that there is a holistic approach to data issues. Over time, such mechanisms can help to change approaches and perceptions of the value that high-quality data holds and can help organizations and whole data ecosystems be more data-driven.

Legal and regulatory frameworks

Legal and regulatory frameworks are crucial to interoperability, especially when it comes to the sharing and integration of data assets between organizations and across national borders. Laws set the boundaries of what is acceptable conduct and what is not. In some instances, they govern *how* data can be shared (for instance, laws that regulate and set standards for data reporting, security and protection) and in others govern *what* data can, or more often cannot, be shared and integrated (for example, data protection and privacy laws).

Laws and regulations exist at many different levels; from the international to the sub-national. International normative frameworks, international laws and domestic laws all set standards for expected conduct and behavior. Meanwhile, memoranda of understanding (MOUs) and various forms of agreement, including data sharing agreements and licenses, set the parameters for specific relationships between organizations. Corporate policies, while not laws, can form part of regulatory frameworks when they set protocols and procedures for data sharing within the parameters of the law. Finally, 'legal interoperability' is itself an important dimension of broader 'interoperability' that relates to how laws from different jurisdictions can be harmonized. Refer to Annex B for further information and a listing of the types of legal mechanism that can support interoperability.

Building a roadmap: an interoperability rapid assessment framework

The following is the first part of the assessment framework produced as part of this Guide. It focuses on the relevance and applicability of conceptual frameworks and the value of institutional frameworks, with a particular focus on legal and regulatory frameworks. It is designed to help inform the development and implementation of data governance strategies and should be supplemented with the resources identified under the Further Reading heading below as well as other context-specific materials.

Action areas	Initial Steps	Advanced Steps
Institutional Frameworks	Identify what model of data governance would work best for your organisation (or you are already a part of) and ensure that interoperability considerations are taken into account from the outset as part of this choice. Put in place a data governance policy that sets out how data is governed across your organisation.	Conduct internal data availability assessments/audits on a regular basis and keep a record of what data is held and handled over its lifecycle. Use this information to periodically review your data governance policy and updated it as required. Conduct comprehensive quality assessments and data audits in collaboration with other stakeholders within the local data ecosystem. Develop Monitoring, Evaluation and Learning frameworks that include indicators on data governance issues.
Oversight and accountability	Identify Data Governance Officers (DGOs) and establish Data Stewardship Teams (DSTs) within your organisation.	Convene Data Governance Councils or Data Governance Steering Committees across organizations comprised of technical, operational and support staff, and supported by the Executive to ensure oversight and accountability.
Legal and Regulatory Frameworks (see Annex B for further information)	Identify and map applicable laws and regulations that apply to the data you hold and process. Identify the types of agreements (MOUs, data sharing agreements, service agreements, licenses, etc.) that are best suited to the organization's needs and adopt templates that can be used by staff to share data, procure IT services, etc. Devise corporate policies that incorporate interoperability-friendly approaches and strategies.	Develop bespoke legal templates for contracts, MOUs and licenses that conform to international best practices and are compatible with other frameworks (for e.g. licenses that are compatible with Creative Commons templates). Where resources permit, provide departmental training and sensitization on how to interpret and implement corporate policies.

Common pitfalls in data governance:

- Failing to take an organisational approach to data management and governance issues and relegating 'data' issues to the IT department;
- Not developing/enforcing a clear chain of accountability specifying roles and responsibilities across departments when it comes to the effective governance of data across/between organisations;
- Overlooking/not considering interoperability issues as a requirement when updating or procuring new IT systems; resulting in internal data silos, and multiple types of data held in incompatible formats and schemas; and
- Not making the best use of legal and regulatory tools and frameworks that can create a safe and structured environment in which data can be shared and integrated while respecting privacy, data protection and security considerations.

Further reading on data management, governance and interoperability

- DAMA International (2017). *Data Management Body of Knowledge*, 2nd ed. New Jersey: Technics Publications.
- Joined-Up Data Standards project (2016). *The frontiers of data interoperability for sustainable development*. Available at: http://devinit.org/wp-content/uploads/2018/02/The-frontiers-of-data-interoperability-for-sustainable-development.pdf
- Palfrey, J. & Gasser, U. (2012). *Interop: The promise and perils of highly interconnected systems*. New York: Basic Books.
- Steiner, R. (2014). *Non-invasive data governance: The path of least resistance and greatest success*. New Jersey: Technics Publications.

Chapter 2: Data and metadata models

Overview

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 or producer of the data set and the date the data was produced – can be organized consistently. Interoperability is highly dependent on data and metadata modelling decisions and practices.

Presently, different organizations, and even different departments within organizations, often handle data and metadata modelling on a case-by-case basis, adopting approaches that are not designed with 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.

An alternative approach is the use of canonical data and metadata models. These are models that follow specific standardized patterns, making them highly reusable and conducive to data sharing. They can be used to represent multiple sources of data and metadata using common patterns, thus making data integration simpler and more efficient.

This chapter makes specific recommendations to overcome structural obstacles to data interoperability, advocating for the use of canonical data models and metadata schemas across systems and organizations. It explores the role of such data and metadata models as enablers of interoperability, explains what data modelling is, and addresses data and metadata standard quality issues. It provides guidance and examples for modelling multi-dimensional data in the development sector, with a view to addressing some of the most common data interoperability challenges. The chapter explores the role of international standards bodies and outlines some widely used metadata schemas, which are particularly suited to ensure the discoverability and usability of collections of data. It also describes how these can help structure data in a way that enhances data interoperability while flexibly accommodating the needs and priorities of different data communities.

The role of data and metadata models as enablers of data interoperability

Both the producers and users of data must have a common understanding of how it is structured in order to effectively exchange it across systems. They must also share a common understanding of how the various components of a dataset relate both to each other and to the components of other datasets. Data and metadata modelling can help to create clarity around these issues and is a critical part of ensuring that systems are designed with interoperability in mind from the outset.

As it stands, a major obstacle to data interoperability in the development sector is the lack of agreement on how to represent data and metadata from different sources in a consistent way when exposing it to

users and third-party applications. Different organizations, or even different departments within an organization, often decide how to structure their data and metadata assets on a case-by-case basis, adopting models of data storage and exchange that respond only to the immediate operational needs of their own processes and applications, without consideration of interoperability needs across the broader data ecosystems they are a part of.

Such "local" or "customized" operational data models become silos because they focus almost exclusively on the efficiency in recording and updating information related to their day-to-day business processes, prioritizing the integrity and accuracy of that information over the analysis needs of external users and the need to integrate their data with external applications and other sources of information. While this is partly an issue of data governance and accountability at the human and institutional layers of interoperability (see the introduction and chapter 1), it also reflects the lack of a coordinated focus on user needs at the technical and operational levels.

FIGURE 8: THE VALUE OF INDUSTRY STANDARDS

Where possible, broadly-accepted and commonly used data models should be adopted by organizations whose data is likely to be shared with others. The models that are developed should not be based on vendor-specific products but on industry standards; those produced by well-respected and accredited standards development organizations. Using standards produced by such bodies, for instance the Unified Modelling Language (UML), produced by the Object Management Group, encourages their widespread adoption and implementation, thus promoting standardization and interoperability across sectors and ecosystems. The benefits of this approach are widely acknowledged in technical circles and numerous bodies exist across sectors that work on establishing consensus, modelling business requirements and data for specific needs and use cases. When organizational commitments to this approach exist, they create an enabling environment that allow interoperability to flourish.

Health Level Seven (HL7), a standards development body is a case in point from the health sector. HL7's Fast Health Information Resources (FHIR) specification provides structure, data elements, data types, and value set information for common components of health data sharing. Similarly, their Clinical Information Modelling Initiative provides detailed clinical health data in common, interoperable models, enabling the exchange of richer health data that can be used for clinical inferencing, including the provision of point-of-care clinical decision support.

Organizations seeking to improve the interoperability of their data should therefore prioritize the needs of client applications and engage with other producers of data in their own sector or domain of operation. They should also engage with key user groups, to ensure consistency of approaches in the selection of data and metadata structures for data sharing and dissemination across organizational boundaries. Key in this endeavor is the adoption of data modelling standards to keep consistent data structures across different databases.

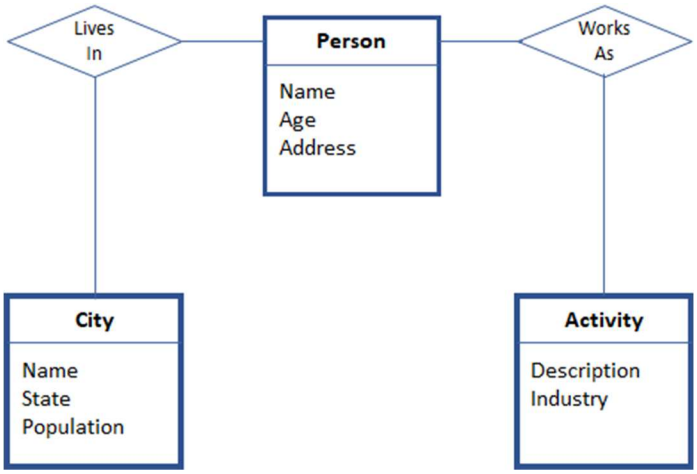
By focusing on interoperability at this level, broader benefits – both in terms of efficiency gains and more consistent and high-quality data – will follow. For instance, common data structures are the foundations that enable developers to more easily create standard application programming interfaces (APIs) to interact with these databases (see Chapter 4).

What is data modelling?

Data modelling is a process focused on clearly and unambiguously identifying things (**entities**) that a dataset aims to capture, and then selecting the key properties (**attributes**) that should be captured to describe those entities in a meaningful way. It requires deciding how entities and attributes relate to each other (**relationships**), and how their information content should be formally codified within a dataset. This is the essence of the Entity-Relationship model, which underlies most modern database management systems and applications.[7]

For example, the content of a dataset may refer to entities such as "city", "person", or "activity", which may be usefully described with attributes like "name", "age", or "industry". And in a specific application, it could be useful to capture the fact that one or more persons may live in a city, that every person has a specific age, and that a person may be employed in one or more types of activity at the same time. Finally, one may require that the names of cities in the dataset be taken from an official list of geographic names, and that the age of a person be represented as a numeric value corresponding to age measured in years.

FIGURE 9: EXAMPLE OF AN ENTITY-RELATIONSHIP MODEL



It is important to set clear expectations from the beginning of the data modelling process regarding the level of detail and the quality of the metadata that will be attached to the data, and to agree on standard naming conventions for all the objects in the model. It is also useful to produce a brief document specifying a governance framework for the data modelling process, including a list of roles and responsibilities and guidelines for sign-off and versioning of data models.

Canonical data and metadata models for interoperability

A data model designed to facilitate data interoperability across systems is different from a model intended, say, to optimize the physical design of a database. The former is focused on simplicity, so the data can be easily understood by a wide range of users and applications, and on being self-contained and stable over time. In contrast, the latter typically emphasizes the elimination of redundancies in data storage, to minimize the cost of maintenance and ensure data integrity.

Canonical models for data exchange and integration are based on fundamental and highly reusable structures that provide a common template of a "user view" to which disparate datasets can be mapped. Thus, instead of each application having to independently understand and individually transform the different internal structures of various data sources, canonical models are used by data providers to expose mixed sources of data and metadata according to common patterns of representation, in this way

[7] See, for instance, Silverston and Agnew (2009).

reducing the number of transformations that user applications need to perform on their own to integrate the data from those sources. In sum, they provide simple templates for data modelling.

This section introduces two canonical models that may be used as a basis for modelling data and metadata elements, so they can be shared in a consistent manner across a wide range of applications. They focus specifically on facilitating the discovery, sharing and use of information by users, as opposed to addressing issues of transaction processing, concurrency control, elimination of redundancies, and data integrity.

The use of canonical data models to support data interoperability requires data providers to take responsibility for implementing any necessary transformations to map the data from its original, operational structures, into commonly agreed presentations for dissemination and distribution purposes (e.g., this may entail the need to undertake so-called "Extract-Transform-Load", or ETL, procedures, hidden from the view of users). The underlying principle is to hide from user the internal complexity of the operational data models (e.g., which are optimized to avoid data redundancy and ensure data consistency validations), so users can concentrate on using data rather than spending time trying to understand the intricacies of internal data structures.

The multi-dimensional 'data cube' model

The **multi-dimensional 'data cube' model** supports data sharing and analysis by presenting information coming from one or more operational databases as a collection of subject-oriented, integrated datasets. This type of model has a long-standing tradition in the business intelligence field, dating back to the 1990s, with the introduction of data warehousing concepts and Online Analytical Processing (OLAP) technologies for enterprise data integration.

A multi-dimensional data cube can be thought of as a model focused on measuring, identifying and describing multiple instances of one type of entity simultaneously. A multidimensional dataset consists of multiple records of **measures** (observed values) organized along a group of **dimensions** (e.g., "time period", "location", "sex" and "age group"). Using this type representation, it is possible to identify each individual data point according to its "position" on a coordinate system defined by a common set of dimensions. In addition to measures and dimensions, the data cube model can also incorporate metadata at the individual data point level in the form of attributes. Attributes provide the information needed to correctly interpret individual observations (e.g., an attribute can specify "percentage" as the unit of measurement).

The definition of each dimension, measure and attribute encapsulates a concept whose domain may or may not be drawn from a code list (for e.g., "country ISO code") or a typed set of values (e.g., "numeric"), or required to adhere to a specific data format (e.g., "YYYY/MM/DD" for dates) or to be contained within a specific range of values (e.g., "numerical values between 0 and 1"). Since in the multidimensional data model the specific values of dimensions and attributes are attached to each individual data observation, each data point can stand alone and be queried and linked with other datasets.

In statistical terms, each data point in a multidimensional dataset can be thought of as a statistical unit belonging to a population. Each statistical unit is characterized by observed values on one or more variables interest, a set of uniquely identifying characteristics, and a set of additional characteristics that further describe it.

The data cube model is not meant to be used as the basis for designing internal or "operational" databases; rather, it is intended to facilitate data exchange and the integration of datasets from disparate sources for analytical purposes. Thus, a multidimensional dataset usually maintains redundant information in a

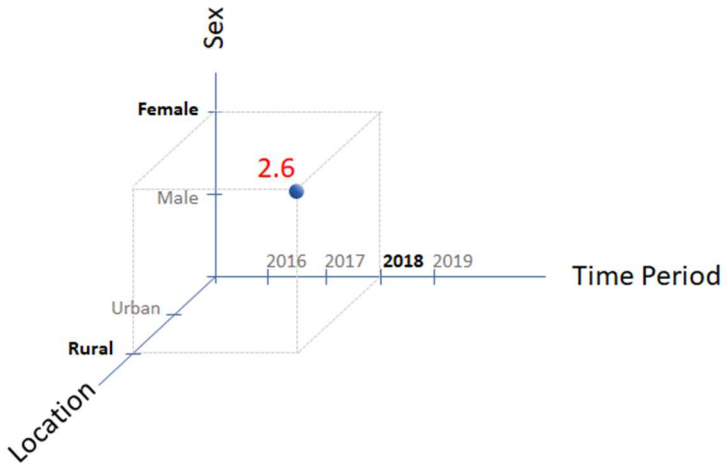
single table (i.e., is not "normalized"), since the intention is to present all relevant data about a population of interest in a simple, self-contained tabular view.

The multidimensional data-cube model can support data interoperability across many different systems, regardless of their technology platform and internal architecture. Moreover, the contents of a multi-dimensional data cube model do not have to be restricted to "small" data sets. In fact, "with the advent of cloud and big data technologies, data-cube infrastructures have become effective instruments to manage Earth observation (EO) resources and services." (Nativi et al 2017).

For example, the data cube representation of an observation on the rate of unemployment for women in rural areas, for the year 2018, can be identified by the following coordinate (dimension) values:

Time period	=	2018
Location	=	Rural areas
Sex	=	Female
Observed value	=	2.6

FIGURE 10: EXAMPLE OF A DATA CUBE MODEL



A subset of observations in a data cube that have a fixed value for all but few dimensions is called a "slice" (e.g., a slice may be a time series consisting, say, of the subset of all observations for a specific location over time). Slices can be used to (1) attach specific reference metadata elements, (2) provide access to subsets of data that are of interest to users or, (3) provide guidance on how to present the data in user applications.

'Slices' of Data Cube

The following table is a data cube representing the annual rate of unemployment by time period, location, and sex. In this example, the columns "Time Period", "Location" and "Sex" are the three dimensions of the data cube. "Unemployment" is the measurement value, while "unit of measurement" provides additional information in the form of a metadata attribute to help interpretation of the data.

Time period	Location	Sex	Unemployment rate	Unit of measurement
2016	Urban	Male	3.4	percent
2016	Urban	Female	3.2	percent
2016	Rural	Male	2.5	percent
2016	Rural	Female	2.3	percent
2017	Urban	Male	3.7	percent
2017	Urban	Female	3.6	percent
2017	Rural	Male	2.7	percent
2017	Rural	Female	2.4	percent
2018	Urban	Male	3.8	percent
2018	Urban	Female	3.6	percent
2018	Rural	Male	2.7	percent
2018	Rural	Female	2.6	percent

A slice representing only urban female unemployment would set the value of the "Sex" dimension equal to "Female" and of the "Location" dimension to "Urban":

Time period	Location	Sex	Unit of measurement
2016	Urban	Female	percent
2017	Urban	Female	percent
2018	Urban	Female	Percent

Dimension values typically have a hierarchical structure that allow users to explore measures at various levels of aggregation/disaggregation. For instance, the "location" dimension may be represented as a hierarchy of municipalities, provinces and countries, whereby the root level of the hierarchy (typically labeled as "all" or "total") groups all observations in the dataset. To ensure correct aggregation of measures, dimension values on the same level of aggregation must be mutually exclusive and exhaustive, and the aggregation function should be such that it is meaningfully applicable. For instance, while it makes sense to calculate the total population of a country by adding population values over all its provinces, it does not make sense to calculate a "total price" by adding the prices over different geographic areas (but one could instead calculate an "average price"). In more general terms, the data cube model is well suited to perform OLAP operations along one or more dimensions, such as roll-up (moving up from a detailed level to a more general level), roll-down (moving down from a more general level to a detailed level), as well as pivoting and filtering.

FIGURE 11: INTEGRATING 'TIME' AND 'SPACE' ACROSS DATASETS

Time and location are two especially important dimensions in any data model. Since they form a part of most datasets (everything takes place somewhere; everything happens at some point in time), time and location are frequently suitable for integration across datasets.

However, time and space are some of the most complex dimensions to model in ways that are interoperable, as they can be represented in multiple ways across different applications and domains. For instance, datasets may vary in their definition of "year" – with some datasets being expressed in "calendar year" and others in "fiscal year", among others. Similarly, changing administrative boundaries and inconsistencies in the naming and definition of geographies at different levels of aggregation is often a challenge for data managers.

The SDMX standard and the multi-dimensional data cube model

Building on a long tradition in the development and implementation of standards for the compilation and exchange of statistical data, the official statistics community increasingly uses the Statistical Data and Metadata Exchange (SDMX) standard to support data exchange and dissemination. SDMX provides a broad set of formal objects to represent statistical data and metadata, as well as actors, processes, and resources within statistical exchanges. The adoption of SDMX as a standard for the exchange of statistical data and metadata has resulted in the increased use of agreed multidimensional data schemas across statistical organizations, enriching them with standard domain-specific and cross-domain concepts and code lists that are made available through central repositories.

The data model underlying SDMX corresponds to the canonical multidimensional data cube model explained above, where every data point has one or more observed values (measures) determined by various identifiers (dimensions) and further described by additional characteristics (attributes). It also provides standardized terminology to name commonly used dimensions and attributes, as well as code lists to populate some of those dimensions and attributes. More specifically, a "Data Structure Definition" (DSD) in SDMX describes the structure of a dataset, by assigning descriptor concepts to the elements of the statistical data, which include:

- Dimensions that form the unique identifier (key) of individual observations;
- Measure(s) which are conventionally associated with the "observation value" (OBS_VALUE) concept; and
- Attributes that provide more information about some part of the dataset.

At the moment, there are various globally agreed DSDs for SDMX in different domains (or sectors) of application, including National Accounts, Balance of Payments, Price Statistics, International Merchandise Trade, Energy, and SDG indicators. SDMX modelling guidelines (SDMX 2018a) and DSD guidelines (SDMX 2018b) provide a step-by-step introduction to data modelling and the creation of DSDs. The guidelines contain numerous links to further, more detailed guidelines and templates that can be used in the modelling process.

Using the Aggregate Data eXchange (ADX) Profile in the Health Sector

The Aggregate Data eXchange (ADX) Profile is an example of a set of standards that is modelled around schemas generated by several different international standards bodies that can also link to national and subnational systems; generating and exchanging data across systems in commonly-used and interoperable data formats. It supports interoperable and routine (weekly, monthly, quarterly, etc.) public health reporting from a health facility or community up-stream to an administrative jurisdiction (for instance a regional health authority). The data that the Profile can exchange can then be used to construct public health indicators, and also contribute to official statistical production at the national level.

ADX can be used to exchange 'data tuples', sets of values captured according to a data element subject, a temporal dimension, and a spatial dimension (i.e. using a simple version of the data cube model outlined above). A hypothetical 'data tuple' would be, for instance, the number of live births recorded in January 2018 at a Nairobi health clinic.

The ADX profile is designed to be interoperable with SDMX and the ADX message structure is defined using a DSD file conformant to the SDMX v2.1 specification. ADX defines a content data structure creator that generates an SDMX v2.1-based DSD and two validation schemas (a W3C XML schema definition (SXD) file and an ISO Schematron) (Schematron 2018) that enable an implementing jurisdiction to formally define the aggregate health data to be exchanged. These 'messages' can then be used by health workers filing routine reports, reporting on health worker data (e.g. number of doctors, nurses, community health workers, etc.), producing monthly summary reports and other compliance documents, as well as global reporting from countries; for example, from a national Health Information Management System (HMIS) to global reporting repositories such as the UNAIDS Global AIDS Response Progress Reporting (GARPR) tool and PEPFAR information system, among others.

FIGURE 12: THE AGGREGATE DATA EXCHANGE-HIV (ADX-HIV) CONTENT PROFILE

Typically, routine HIV reports are submitted from health facilities to an administrative jurisdiction such as a health district or health department, and eventually to the national level where indicator data are aggregated for global reporting on the HIV/AIDS response. The ADX-HIV Content Profile uses the ADX Profile to describe core indicators within data elements and associated disaggregated data and facilitates the generation and exchange of ADX conformant messages for aggregated HIV data.

The HIV core indicators used in the Profile are drawn from indicator sets published and used across Ministries of Health, the World Health Organisation (WHO), the Global Fund, UNAIDS and others. The Profile focuses on indicators used for monitoring progress towards achieving HIV epidemic control and the 90-90-90 Global Goals. These Goals, to be accomplished by 2020 are: for 90% of people with HIV to be diagnosed and know their status; for 90% of people living with HIV to be accessing and using Antiretroviral Therapy (ART); and, for 90% of people on ART to be virally supressed.

The data elements for the HIV core indicators are used to generate a set of artefacts including a common DSD, as well as the XML schema and schematron mentioned above thus enabling the system to leverage ISO Standard 8601 to codify age groups, HL7 administrative sex to codify sex, and SNOMED-CT to codify the results of HIV tests. Producing the data in this way facilitates the routine reporting of large quantities of data from health facilities to global repositories and is used for global reporting by countries.

This is an example of how different interoperable standards and schemas can be used in conjunction to create systems that are interoperable across different levels within specific sectors/domains.

Standard metadata schemas

A metadata schema specifies the metadata elements that should accompany a dataset within a domain of application. For instance, W3C's Data Cube Vocabulary (W3C 2014) recommends that each dataset should be accompanied by title, description, date of issue or modification, subject, publisher, license, keywords, etc.

The Data Catalog (DCAT) vocabulary

The Data Catalog (DCAT) Vocabulary is a canonical metadata schema specification, designed to publish metadata in a human and machine-readable format and to improve discoverability of datasets, support dataset management and address operational issues, while promoting the re-use of metadata elements from existing namespaces. It is a well-documented, flexible and practical metadata standard grounded on

the foundations of Dublin Core[8], SKOS (Simple Knowledge Organization System)[9] and FOAF (Friend of a Friend) vocabularies[10], which opens the door to cross-map different DCAT implementations to one another (Lisowska 2016). Moreover, DCAT is used by major engines for data portals, such as CKAN, DKAN, Socrata and OpenDataSoft. And has been adopted by the European Commission to publish datasets pertaining to the public sector in Europe.[11]

DCAT is based on the three main classes of data objects, namely: "Data Catalog", "Data Set", and "Distribution". While the multidimensional data model allows to model metadata at the observation level, the DCAT metadata schema is used to organize reference metadata at the level of the dataset and above. For instance, the metadata elements recommended to describe a "Data Catalog" include information such as its description, language, license, publisher, release date, rights, etc. The recommended metadata elements of a "Dataset" include items such as contact point, description, distribution, frequency, identifier, keywords, language, publisher, release date, title, update date, etc. Finally, the recommended metadata elements of a "Distribution" object include information on description, access URL, byte size, download URL, format, license, media type, release date, rights, title, and update date. The table below provides a description of the main metadata elements that are part of the DCAT model.

One key feature of the DCAT vocabulary is its extensibility. It can be adapted to meet the needs of different groups of data publishers through so-called "application profiles". For instance, the DCAT-AP v1.1 introduced the new "Data Portal" element to represent a specific web-based system that contains a data catalog with descriptions of data sets, and which provides services that enable users to discover data content. It also introduced new properties for existing metadata objects. Moreover, two specific application profiles of DCAT (GeoDCAT[12] for geospatial information and StatDCAT[13] for statistical data) which are interoperable with Geographic Information Metadata standard ISO19115[14] and with SDMX, respectively, are currently being developed.

Quality of data and metadata models

Data can be represented using different schemas (e.g., relational, multi-dimensional, object-oriented, etc.). The decision as to what kind of schema to use to best describe a dataset will depend on the specific business requirements and the specific domain of application. However, the quality of any data model should be assessed in relation to some general criteria, which include, among others:

- **completeness**, or the extent to which all relevant entities and relations are captured by the model;
- **simplicity**, or the ability to provide a concise representation of all key aspects of the data with few entities, relationships, and attributes, and with a small number of instances thereof;
- **standardization**, or the use of generic or canonical structures (such as commonly definition of entities such as "person" or "address") and re-usable code lists to specify concept domains;

[8] For further information see: http://dublincore.org/documents/dcmi-terms/
[9] For further information see: https://www.w3.org/2004/02/skos/
[10] For further information see: http://xmlns.com/foaf/spec/
[11] See, for instance, https://www.europeandataportal.eu/en. More information on the DCAT Application Profile for data portals in Europe is available from https://ec.europa.eu/isa2/solutions/dcat-application-profile-data-portals-europe_en
[12] For further information see: https://joinup.ec.europa.eu/release/geodcat-ap/v101
[13] For further information see: https://joinup.ec.europa.eu/solution/statdcat-application-profile-data-portals-europe
[14] For further information see: https://www.iso.org/standard/53798.html

- **flexibility**, or the ability to easily add new entities and attributes without compromising the integrity of existing data and applications;
- **readability**, or general adherence to naming conventions, as well as overall quality of definitions.

Building a roadmap: an interoperability rapid assessment framework

The following is the first part of the assessment framework produced as part of this Guide. It focuses on the relevance and applicability of conceptual frameworks and the value of institutional frameworks, with a particular focus on legal and regulatory frameworks. It is designed to help inform the development and implementation of data governance strategies and should be supplemented with the resources identified under the 'Further Reading' heading below as well as other context-specific materials.

Action areas	Initial Steps	Advanced Steps
Modelling data structures	Starting from a set of source tables, identify elementary datasets to be modelled (variables or indicators). Identify key entities that are described in the information contained in the dataset (e.g., places, people, businesses…): • Identify the dimensions and attributes needed to describe each entity at the target level of granularity (e.g., location, time period, sex…); • To the extent possible, re-use standard dimensions and naming conventions from existing data models (e.g., from existing SDMX data structure definitions); • Consider merging or splitting columns from original tables to define more useful dimensions for data exchange. Create a separate table of distinct values for each dimension, assigning a unique numeric ID to each row.	Identify or define relevant hierarchies of broader/narrower (parent/child) levels for the dimensions of the dataset. Consider enriching the set of attribute columns with information from other datasets, with a view to make each record self-contained and self-explanatory.
Modelling metadata	Identify a minimum set of metadata elements relevant to describe the dataset.	Map all relevant metadata elements to DCAT vocabulary classes.

Quality considerations	Internally consider user needs and data quality considerations when deciding an approach to modelling.	Establish feedback loops with user groups to enable them to comment on the usability and usefulness of the models that have been employed. Utilise the feedback that is received to regularly update the system and maintain high data quality and usability standards.
Common pitfalls in data modelling: • A common mistake in modelling datasets for sharing and dissemination is to try to replicate internal data structures from operational database systems. The focus should be on producing simple, self-contained datasets that are easy to understand and manipulate by users and client applications.		

Further reading on data and metadata modelling

- Geppert, A. (2018). *Data Warehouse Design & Multi-dimensional Models*. Available from: https://www.ifi.uzh.ch/dbtg/teaching/courses/DataWarehousing/Unterlagen/dwh-04.pdf
- Silverston, L. and Agnew, P. (2009). *The Data Model Resource Book, Volume 3: Universal Patterns for Data Modeling*. New York: Wiley Publishing.

Data modelling tools

- **BigGorilla:** is an open-source platform for data integration in Python. It offers components for data integration and data preparation that can be combined and reused in different contexts and different ways (https://www.biggorilla.org/).
- **DSD Constructor**: Intuitive desktop tool for creating data structure definitions to map and build tables. Complements the ILO Smart inputs (https://www.ilo.org/ilostat/tools/dsdConstructor/Install.htm).
- **ILO Smart tool:** The Statistical Metadata-driven Analysis and Reporting Tool is a statistical processor and transcoding tool able to produce datasets processing microdata or aggregate data in several formats, according to the structural metadata read from a Dataflow or DSD. Output files can be generated in diverse formats, intended for analysis, data reporting or to feed a dissemination platform. A very useful tool when producing SDMX datasets for the SDGs (or any) data reporting (https://www.ilo.org/ilostat/tools/smart/index.html).
- **Joined-up Data Standards Navigator spreadsheet tool:** A Google Sheet Add-on facilitates matching fields in one data standard to those in another http://joinedupdata.org/en/spreadsheet.
- **OpenRefine:** for working with messy data and easily identifying problems http://openrefine.org/.
- **SDMX reference infrastructure:** A suite of tools that includes support for data mapping https://sdmx.org/?page_id=4666.
- **The Data Structure Wizard:** is a desktop application that is able to convert/edit commonly used metadata formats into SDMX-ML formats. It contains an interface that allows the user to select a given Data Structure and to complete the data according to requirements: https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Data_Structure_Wizard_DSW.

Chapter 3: Standard classifications and vocabularies

"A good classification functions in much the same way that a theory does, connecting concepts in a useful structure." (Kwasnik 2000)

Overview

The previous chapter introduced standard data models and metadata schemas as key facilitators of data interoperability that can be used to standardize the way in which datasets are structured. This chapter provides a more granular view of how standard classifications and vocabularies enable semantic interoperability across different applications and 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 allows data to be shared efficiently and enables users to more easily find related information across 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 of the International Council for Science (CODATA) 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.

To expose data without ambiguities and ensure semantic interoperability, it is crucial to focus on the adoption of standard vocabularies and classifications early on, starting at the design phase of any new data collection, processing or dissemination system. Simultaneously with the data modelling decisions, it is important to identify any relevant, publicly available, and widely used classifications and vocabularies that could be re-used to codify and populate the content of dimensions, attributes, and measures in the data set, as well as any additional metadata elements used to describe it and catalogue it.

It is important to acknowledge, however, that the use of customized classifications and vocabularies is sometimes unavoidable, either because they are part of legacy data management systems, or because of very specific needs in their primary domain of application. In those cases, it is important to engage in structural and semantic harmonization efforts. For instance, it is often necessary to "standardize after the fact" (McKeever and Johnson 2015), mapping "local" terminology used to designate measures and dimensions to commonly used, standard vocabularies and taxonomies. Similarly, any custom hierarchies used to group observations at different levels of aggregation should be mapped to standard classifications (see, Hoffman and Chamie 1999).

This chapter explores the role of standard classifications and vocabularies in data interoperability, highlighting examples of commonly used classifications and vocabularies in the development sector and the need for governance of standard classifications and vocabularies.

Role of standard classifications and vocabularies in data interoperability

The choice of terms used to describe the elements of a dataset can have a significant impact upon the interoperability and overall usability of the dataset. Classifications and vocabularies both enable and constrain the way in which data is collected, organized, analyzed, and disseminated, and help harmonize the process of knowledge creation, knowledge management, and knowledge sharing. In short, they allow

"data to be cleanly comparable and aggregates to be computable" (Srinivasan 2017). Therefore, the use of standard classifications and vocabularies to identify, label, and catalogue individual data points and datasets has an impact on the ability of people (and machines) to easily find, access and integrate different datasets.

Standard vocabularies and classifications help improve consistency in the description of data and metadata elements within a dataset. They allow data producers to express the meaning of data without ambiguities and enable users to find and link related pieces of information, from the unit record level to the dataset level, across different information systems. Although not all metadata elements lend themselves to classification or vocabulary control (e.g. titles are usually free-text metadata elements), the use of standard classifications and vocabularies to constrain the set of values that populate dimensions and attributes of a dataset, as well as key reference metadata elements, contributes to the interoperability and harmonization of data from different sources. For instance, the adoption of common classifications to model the hierarchical levels of common data-cube dimensions (see chapter 2) across different datasets offers the ability to combine and correlate them and to perform joint queries across them (Torlone 2009).

To meet the needs of a continuously changing data ecosystem, classifications and vocabularies need to adapt over time and be continuously "mapped" to each other by establishing associations of correspondence between their elements. Moreover, they need to be publicly available and accessible in standard formats, such as CSV, JSON or RDF (see Chapter 5 for more details on the application of RDF).

FIGURE 13: THE JOINED-UP DATA STANDARDS NAVIGATOR

The Joined-up Data Standards Navigator is a network of mapped data standards or classifications of data that are relevant to the field of development. The network contains mapped standards and classifications covering socioeconomic sectors, indicators, country classifications and surveys. The Navigator was developed as a freely available resource for those needing to work across multiple data standards, removing the need for manual work to understand how one standard or sector code relates or compares to another. It can benefit users who work with data relating to development either at an international or a national level.

The Navigator cross-links standards in a machine-readable way to enable users to understand the relationships between different international data standards, and to make comparisons between them. The data standards held in the Navigator network are linked together using the Simple Knowledge Organization System (SKOS). This provides a universal and established 'language' for defining relationships between concepts (terms) and also provides a means of systematically comparing concepts across the mapped data standards. Using SKOS allows the standards to be connected by not only one-to-one linear relationship but also defines more complex relationships such as one-to-many or many-to-one. SKOS can also describe if concepts in comparable standards can be used interchangeably (are exactly the same) or not (they can be closely linked).

The relationships between standards can be retrieved either by browsing the contents of the Navigator's projects or by using available tools to explore if and how the data standards compare. The 'spreadsheet tool' is a Google sheet Add-on that allows the user to discover mappings within their working spreadsheet; the 'search and translate tool' on the other hand allows for both simple and advanced search Navigator's content online.

Controlled vocabularies

Controlled vocabularies are restricted lists of terms that can be used for indexing, labelling and categorizing the content of information resources. They help ensure consistency and avoid ambiguity in the description of data. They are also subject to specific policies that determine who may add terms to

the list, when and how, and usually contain guidance regarding the use of "preferred" over "non-preferred" terms to describe specific concepts (see, Hedden 2010).

There are different types of controlled vocabularies, depending on the relationships between the terms that constitute them. For instance, controlled vocabularies that include additional information about the usage of each term and about its relationship to broader, narrower, related and or equivalent terms, are also known as "thesauri", and they are an indispensable tool for making data content searchable and findable. Development Initiatives' Joined-Up Data Standards Navigator described in Figure 13 is an example of a thesaurus that powers semantic interoperability tools and solutions. Another example is the UNESCO Thesaurus[15], consisting of a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in the fields of education, culture, natural sciences, social and human sciences, communication and information.

Standard classifications

Standard classifications, like controlled vocabularies, help data producers to organize complex sets of information according to semantic categories. However, they provide more structure than controlled vocabularies, as their terms correspond to a strictly exhaustive list of mutually exclusive categories, often presented in hierarchical form, which can be used to unambiguously categorize all the objects that fall within a specific domain. For example, the International Classification of Diseases[16] (ICD) maintained by the WHO consists of a limited number of mutually exclusive categories that encompass the complete range of morbid conditions. This and other international standard classifications allow developers of database systems and data dissemination platforms to use standardized, widely accepted codes to represent the values for specific dimensions, attributes or even measures of a dataset.

There are nevertheless limits to the amount of information that can be embedded in the hierarchical structure of a classification before it becomes too complex (Kwasnik 2000). In practice, it is impossible to capture in a single classification all aspects of reality that are relevant to a particular domain, and the use

FIGURE 14: HARMONIZED COMMODITY DESCRIPTION AND CODING SYSTEMS

The Harmonized System (HS) is an international nomenclature for the classification of products. It allows countries to classify traded goods consistently for customs purposes. At the international level, the HS for classifying goods utilizes a six-digit coding system. The HS comprises approximately 5,300 article/product descriptions that appear as headings and subheadings, arranged in 99 chapters, grouped into 21 sections. The six digits can be broken down into three parts. The first two digits (HS-2) identify the chapter the goods are classified in, e.g. 09 = Coffee, Tea, Maté and Spices. The next two digits (HS-4) identify groupings within that chapter, e.g. 09.02 = Tea, whether or not flavored. The next two digits (HS-6) are even more specific, e.g. 09.02.10 Green tea (not fermented).

The Harmonized System was introduced in 1988 and has been adopted by most countries worldwide. It has undergone several changes in the classification of products. These changes are called revisions and entered into force in 1996, 2002, 2007, 2012 and 2017. The amendments (split, merge, change in scope) between the latest HS edition and its previous edition are maintained by the World Customs Organization. The UNSD then harmonizes those relationships in to 1-to-1, 1-to-n, n-to-1 and n-to-1 *mappings*, and extends the correspondences to earlier HS editions, as well as to other standard classifications, such as the Standard International Trade Classification and the Classification by Broad Economic Categories. The information and its methodological papers are available for download from UNSD Commodity Correspondence Tables.

[15]For further information see: http://vocabularies.unesco.org/browser/thesaurus/en/
[16]For further information see: http://www.who.int/classifications/icd/en/

of different, possibly overlapping hierarchies may be needed in order to expose a dataset to different types of users.

Common classifications and vocabularies in the development sector

Researchers and practitioners in many fields, from official statistics to library and information science, have a long-standing tradition in the development of standard classifications and vocabularies. Examples that are used by a range of stakeholders within the development sector include:

- The SDMX content-oriented guidelines (SDMX 2018c) include a Glossary with concepts and related definitions used in structural and reference metadata by international organizations and national data-producing agencies, as well as a set of various code lists and a classification of subject-matter domains. The content-oriented guidelines define a set of common statistical concepts and associated code lists that con be re-used across data sets. Data providers are encouraged to make direct reference to terms in the SDMX glossary to improve interoperability and comparability across datasets. Some of the most common concepts included in the SDMX content-oriented guidelines, which correspond to the data-cube model, include the time dimension ('REF_PERIOD'), the geographic area dimension ('REF_AREA'), measure ('OBS_VALUE') and unit of measurement ('UNIT_MEASURE').
- Commonly used geographic referencing and coding standards include the standard country or area codes for statistical use (M49)[17], as well as the ISO 3166[18] standard for country codes and codes for their subdivisions.
- Dublin Core[19] is a standard vocabulary widely used to represent key metadata annotations in many fields. Dublin Core Terms should be used as much as possible to represent commonly needed structural and reference metadata elements.

The governance of standard classifications and vocabularies

Classification systems should be used, "with full comprehension of the meaning of each code" (Vancauwenbergh 2017). However, standard classifications are often published using only codes and related terms, without semantic definitions and explanatory notes. This opens the door to different interpretations of the same terms, which can result in confusion and, at a systemic level, a situation where different organizations are essentially defining the same terms differently within their data models, harming their potential for interoperability.

While the design of interoperable information systems requires users to structure and present data according to information categories that are meaningful to multiple target audiences, no designer can guarantee that his or her intended attributions of meaning will be universally accepted.

Controlled vocabularies and classifications need to have governance mechanisms and policies in place to determine how they are maintained and by whom. They also need to be flexible and be continuously updated and adapted to the diverse and changing interests of data producers, data managers, and data users. But the process of producing and updating standard classification systems and vocabularies is not only technical; it requires the active participation of multiple stakeholder groups. The starting point should always be to review classifications and vocabularies that are already in use and, to the degree possible,

[17]For further information see: https://unstats.un.org/unsd/methodology/m49/
[18]For further information see: https://www.iso.org/iso-3166-country-codes.html
[19]For further information see: http://dublincore.org/documents/dcmi-terms/

re-use them. New classifications and vocabularies should only be suggested where a genuine gap exists or where old systems are redundant and no longer fit for purpose.

Once a standard classification or vocabulary is created and implemented; policies, methods and procedures for its maintenance, including rules for adding, changing, moving or deleting terms or relationships, as well as the identification of roles and responsibilities need to be produced. These components should form part of organizations' data governance strategies and should also be subject to oversight by Data Governance Officers and Data Governance Councils/Committees (see Chapter 1).

Building a roadmap: an interoperability rapid assessment framework

The following is the first part of the assessment framework produced as part of this Guide. It focuses on the relevance and applicability of conceptual frameworks and the value of institutional frameworks, with a particular focus on legal and regulatory frameworks. It is designed to help inform the development and implementation of data governance strategies and should be supplemented with the resources identified under the Further Reading heading below as well as other context-specific materials.

Action areas	Initial Steps	Advanced Steps
Using common classifications and vocabularies	Identify relevant, publicly available, and widely used classifications and vocabularies that can be re-used to codify and populate the content of dimensions, attributes, and measures in a data set. Adopt standard vocabularies and classifications early on, starting at the design phase of any new data collection, processing or dissemination system.	Collaborate with members of other data communities in the development and use of common classifications and vocabularies to describe cross-cutting concepts and relationships. Make any new classifications or vocabularies publicly available and accessible in standard formats, such as CSV, JSON or RDF.
Creating semantic interoperability between classifications	Engage in structural and semantic harmonization efforts, mapping "local" terminology used to designate measures and dimensions to commonly used, standard vocabularies and taxonomies.	Map any custom hierarchies used to group observations at different levels of aggregation to standard classifications.
Governance considerations	Establish policies, methods and procedures to maintain classifications and vocabularies, including rules for adding, changing, moving or deleting terms or relationships, as well as the identification of roles and responsibilities.	Seek the active participation of multiple stakeholder groups in the process of producing and updating standard classification systems and vocabularies.
Common pitfalls when using classifications and vocabularies:		

- As information systems become larger, more complex and more interconnected, there is a growing tension between the need to use standard classifications and vocabularies to enhance interoperability, and the need to devise specialized ones for specific user groups. Thought needs to be put into how this balance is set and to the extent possible, efforts should be made to connect new classifications to existing ones to ensure continuity and interoperability down the line.
- Classifications and controlled vocabularies should not be seen as static; they need to be flexible and be continuously updated and adapted to the diverse and changing interests of data producers, data managers, and data users.

Further reading on standard classifications and vocabularies

- Hedden, H. (2016). *The Accidental Taxonomist*. Medford: New Jersey.
- Hoffman, E. and Chamie, M. (1999). *Standard Statistical Classification: Basic Principles.* United Nations: New York. Available at: https://unstats.un.org/unsd/classifications/bestpractices/basicprinciples_1999.pdf
- Kwasnik, B. H. (2000). 'The role of classification in knowledge representation and discovery', *School of Information Studies: Faculty Scholarship*, 147. Available at: https://surface.syr.edu/istpub/147
- McKeever, S. and Johnson, D. (2015). 'The role of markup for enabling interoperability in health informatics', *Frontiers in Physiology*, 6, pp.1-10.
- Srinivasan, R. (2017). *Whose global village? Rethinking how technology shapes our world.* NYU Press: New York.
- Torlone, R., (2009). Interoperability in Data Warehouses, *Encyclopedia of Database Systems*. Available at: http://torlone.dia.uniroma3.it/
- UK National Statistics (2008). *National Statistics Code of Practice: Protocol on Statistical Integration and Classification*. Revised version 1.2, Available at: https://www.ons.gov.uk/ons/guide-method/the-national-statistics-standard/code-of-practice/protocols/statistical-integration-and-classification.pdf
- Vancauwenbergh, S. (2017). 'Governance of Research Information and Classifications, Key Assets to Interoperability of CRIS Systems in Inter-organizational Contexts', *Procedia Computer Science*, 106, pp.335–342.

Chapter 4: Open data formats and standard interfaces

Overview

The adoption of common data and metadata models (chapter 2), and the use of controlled vocabularies and classifications to standardize their content (chapter 3), are necessary, but insufficient conditions to achieve data interoperability. In addition to the considerations and steps set out in chapters 2 and 3, standardized datasets need to be expressed in formats and made available through means that enable both machine-to-human and machine-to-machine access and use. In other words, once the basic structure of a dataset is in place and its contents are codified using standard classifications and controlled vocabularies, the data then needs to be made easily available and accessible to a variety of user groups.

Interoperability is therefore not only about standardized data production, but also about standardized "data logistics" (Walsh and Pollock n.d.), meaning that it requires the use of common patterns and pathways to get data from providers to users in a fast, convenient, effective, and efficient manner.

This section provides an overview of various approaches that exist to make data discoverable and present it so that developers and end-users can access data in more reliable and straightforward ways. The chapter recommends a set of data formats (among many others that exist) and the development of application programming interfaces (APIs) and user interfaces to support interoperability.

Open data formats

Electronic data files can be created in many ways, and data interoperability is greatly enhanced if data is made available using openly documented, non-proprietary formats. For maximum interoperability, data and metadata files need to be published in human-editable and machine-usable ways, and need to be agnostic to language, technology and infrastructure. A first step is to make the data available through bulk downloads in open data formats. There are various fully documented and widely agreed-upon patterns for the construction of digital data files, such as CSV, JSON, XML, and GeoJSON, among many others.

CSV is an easy-to-use data format for both developers and non-developers alike. The CSV serialization format is probably the most widely supported across different technological platforms, and although it does not incorporate a schema for validation, there are recent alternatives that combine CSV tabular data with additional schema information in containerized data packages (see, for instance, the Frictionless Data Packages described in Figure 15 below).

The use of JSON or XML to structure data in a text format and to exchange data over the internet is particularly useful for data interoperability, since these serializations allow producers and users to encode common data elements and sub-elements in such a way that data and metadata are linked together but clearly distinguishable from each other. It is important to note that while XML and JSON offer a common syntactic format for sharing data among data sources, they alone cannot address semantic integration issues, since it is still possible to share XML or JSON files whose tags are "completely meaningless outside a particular domain of application." (Halevy and Ordille 2006).

FIGURE 15: THE DATA PACKAGE STANDARD

The Data Package standard is a containerization format used to describe and package any collection of data, based on existing practices for publishing open-source software. It provides a "contract for data interoperability" through the specification of a simple wrapper and basic structure for data sharing, integration and automation without requiring major changes to the underlying data being packaged.

Its specification is based on the principles of simplicity and extensibility, and the ability to provide both human-editable and machine-useable metadata. It emphasizes the re-use of existing standard formats for data, and is language, technology and infrastructure-agnostic. To create a data package, one has only to create a 'datapackage.json' descriptor file in the top-level directory of a set of data files. This descriptor file contains general metadata (e.g., name of the package, licence, publisher, etc.) as well as a list of all the other files included in the package, along with information about them (e.g., size and schema).

Data serializations in SDMX

SDMX is a data standard that encompasses a data model (the multidimensional data cube), standard vocabularies (content-oriented guidelines), a formal schema definition (DSD), and various data serialization formats for the construction of data files and electronic messages for data exchange.

In the context of SDMX, data providers can choose between data serialization formats for sharing datasets, including XML, CSV, JSON or even EDIFACT[20]. "The SDMX Roadmap 2020 foresees the promotion of easy-to-use SDMX-compatible file formats such as CSV. The most important thing about these formats is that, despite their compact size, the data structure defined by the SDMX metadata is still complied with." (Stahl and Staab 2018, p. 97).

Application programming interfaces

Client applications help users discover, access, integrate and use data from multiple sources. Once a standard serialization format is in place, data providers can think of more sophisticated ways of making the data available, such as the use of APIs to deliver data resources over the web to multiple groups of users.

APIs are highly-reusable pieces of software that enable multiple applications to interact with an information system. They provide machine-to-machine access to data services and provide a standardized means of handling security and errors. When APIs behave in predictable ways, accept requests from users following well-known syntax rules, and yield results that are easy to understand and to act upon, it is possible to automate data flows that involve repetitive and frequent data sharing and exchange operations, avoiding costly and error-prone manual intervention. This allows users to focus on the data rather than spend their time collecting it. APIs provide the building blocks for users to easily pull the data elements they need to build their applications. In a sense, APIs are the virtual highways that allow data to travel back and forth between different websites and platforms.

API documentation is a technical contract between a data provider and its users. As such, it should describe all the options and resources that are available, as well as the parameters that need to be provided by the client to the server, and the content, format and structure of the resulting information that is sent back to the calling application, including error messages and sample responses. One good

[20] For further information see: https://www.unece.org/cefact/edifact/welcome.html

practice in the development of APIs is the consistent use an API description language (e.g., Swagger[21]) to document their functionality. It is also crucial for the documentation of an API to keep track of its different versions (Biehl 2016).

Web APIs

A web service delivers the result of a data processing task performed by a specialized computer (the server) upon request by other computers. Web APIs serve information resources to client applications through the internet, thus enabling the emergence of modern distributed data processing and analysis systems with loosely coupled components. They enable the implement of a service-oriented architecture where information resources are provided to data users upon request, independently and with no need for prior knowledge of the specifications that are used by the requesting applications.

The use of common standards and design patterns in the implementation of web APIs allows multiple developers to easily consume and integrate data resources over the web, opening up new possibilities for on-the-fly generation of data analysis and visualizations. The adoption of web APIs for the dissemination of data, based on open specifications, is a crucial step towards improved data interoperability.

FIGURE 16: THE OPENAPI SPECIFICATION

The OpenAPI Specification (OAS, formerly known as the Swagger Specification), "defines a standard, language-agnostic interface to RESTful APIs which allows both humans and computers to discover and understand the capabilities of the service", allowing users to "understand and interact with the remote service with a minimal amount of implementation logic."

In simpler terms, it establishes a standard format to document all the functionality of a web REST API, describing, in a way that is both human and machine readable, the resources it provides and the operations that can be performed on each of these resources, as well as the inputs needed for, and outputs provided by, each of these operations. It also can be used to document user authentication methods, and to provide additional information like contact information, license, terms of use, etc.

In the context of microservice and service-oriented architectures, the OpenAPI Specification has emerged as the standard format for defining the contract between client applications and services exposed via APIs, making it easier to orchestrate applications as collections of loosely coupled services, each of which supports self-contained business functions (Vasudevan 2017).

One example of a web API that provides access to data on SDG indicators following the OpenAPI specification is the UNSD'S Global Sustainable Development Goal Indicators API. This API enables developers to use indicator data in a flexible manner directly within their own applications. This dramatically lowers data maintenance costs and ensures their applications always contain official and up-to-date indicator data, straight from the source.

API interoperability

The explosive growth in availability of open APIs within and across organizations has led to a new interoperability challenge, namely that of integrating multiple APIs. Whereas building integrated systems from scratch would be extremely costly and disruptive, one approach to deal with legacy systems that do not "talk to each other" is therefore to build a middleware layer that connects one or more client applications with the legacy data systems through common API specifications (Feld and Stoddard 2004).

[21] For further information see: https://swagger.io/specification/

The use of standardized APIs across different data platforms allows application developers to quickly "mash up" data from multiple sources. Such APIs function as "middle tier" lenses that allow developers to view individual data assets as building blocks for their applications, which they can then put together in different combinations to address specific user needs. APIs should be designed with the needs of application developers in mind, focusing on helping them create information products that satisfy the requirements of end users. In this context, APIs need to be well-documented, easy to understand, and easy to integrate with other systems.

FIGURE 17: BUILDING EFFECTIVE APIS

As part of its API Highways initiative, the GPSDD has developed an API Playbook, a concise and accessible guide to build web services that can be seamlessly integrated by the global developer community. The API Playbook includes checklists and key questions drawn from successful practices from the private sector and government to help build better digital services in support of the SDGs. It is organized around the following 8 "plays" or recommendations:

1. Document everything;
2. Consider the Developer Experience;
3. Be an upstanding citizen of the web;
4. Be a part of the community and ask for help;
5. Build trust;
6. Consider the future;
7. Plan for the long tail;
8. Make it easy to use.

Broadly speaking, it is good practice to manage all of an organization's APIs as a single product. Efforts should be made to ensure that all component APIs are standardized and have mutually consistent documentation and functionality and implement common design patterns built from reusable components. Moreover, adopting a common set of basic functionalities and design patterns is crucial to improve interoperability across different APIs. For instance, the error messages from all APIs in an organization's API portfolio should have the same structure. To facilitate discoverability, the description of the API portfolio should be served by a specific end. This enables the caller to discover each API within the portfolio by downloading and parting the API portfolio.

However, there are often cases in which organizations need to maintain multiple APIs which are not fully standardized, say, because they are legacy services developed in isolation, or because they need to deviate from common patterns to deliver customized services to highly specialized client applications. In those cases, a useful approach to improve interoperability of APIs is the creation of API aggregation services, or "API mash ups". This are services that draw data from multiple APIs (from one or more providers) and repackage it to create new, integrated data services for end users. Such API aggregation services can greatly improve developer experience and generate value by eliminating the need to work with multiple APIs. This is the approach followed, for instance, by the API Highways initiative described in Figure 18.

Standardized user experience

The user interface of a data dissemination platform is the point at which most lay users first come into contact with data. Standardizing this experience can help promote data use and usability. Interfaces should therefore be designed to behave consistently and in familiar ways, following common design patterns and rules of communication, so users can easily and intuitively interact with them instead of having to invest large amounts of time and effort trying to understand the rules of engagement whenever they are confronted with a new data source.

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 interoperate and 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 needs.

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. Similarly, using standardized patterns and reusable building blocks when designing human-machine interfaces across different applications can significantly reduce the effort that users have to invest to find and use the data they need.

Building a roadmap: an interoperability rapid assessment

The following is the first part of the assessment framework produced as part of this Guide. It focuses on the relevance and applicability of conceptual frameworks and the value of institutional frameworks, with a particular focus on legal and regulatory frameworks. It is designed to help inform the development and implementation of data governance strategies and should be supplemented with the resources identified under the Further Reading heading below as well as other context-specific materials.

Action areas	Initial Steps	Advanced Steps
Using open data formats	Make the data available through bulk downloads, for example as CSV files.	Use XML, JSON, GeoJSON or other widely-available open data formats to encode common data elements and sub-elements in such a way that data and metadata are linked together but clearly distinguishable from each other. Use a data containerization format such as the Data Package standard format to publish data sets.
Using standard APIs	Set up a webpage and document all the functionality of existing web APIs in use, describing the resources being provided, the operations that can be performed, as well as the inputs needed for, and outputs provided by, each of operation. Provide additional information such as contact information, any licences used, terms of use, etc.	Develop RESTful web APIs for data dissemination following the OpenAPI specification and document them fully using an API documentation language such as Swagger. Expose datasets using SDMX web services. Serve the description of the API portfolio by a specific end point, to facilitate the discoverability and use of APIs.
Enhancing user experience	Follow common design patterns and rules of communication, so users can easily and intuitively interact with system interfaces.	Create or subscribe to API aggregation services, or “API mashups” to improve developer experience and generate value by eliminating the need to work with multiple APIs.

Common pitfalls when using open data formats and standardized interfaces:

- Over-customization of an interface can inhibit its accessibility, usability and interoperability with other systems. System interfaces should prioritize interoperability and flexibility over specificity and optimization. In short, a balance must be struck between specific user group needs and broader usability.
- Notwithstanding the above, the ability to customize and alter interfaces for specific use cases should be planned for as part of a broader data governance policy.
- All the APIs of an organization should be managed as one product, making sure that all component APIs are standardized and have mutually consistent documentation and functionality, and implement common design patterns built from reusable components.

Further reading on open formats and standard interfaces

- Biehl, M. (2016). *RESTful API Design: Best Practices in API Design with REST*. API University Press.
- Feld, C. S. and Stoddard, D. B. (2004). Getting IT Right, *Harvard Business Review*. Available at: https://hbr.org/2004/02/getting-it-right
- Halevy, A. et al. (2017). Data Integration: After the Teenage Years, *in* Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. Available at: https://dl.acm.org/citation.cfm?id=3034786
- Stahl, R. and Staab, P. (2018). *Measuring the Data Universe*. New York: Springer Press.
- Vasudevan, K. (2017). *Microservices, APIs, and Swagger: How They Fit Together.* Available at: https://swagger.io/blog/api-strategy/microservices-apis-and-swagger
- Walsh, P. and Pollock, R., (No date). *Data Package Specifications.* Available at: https://frictionlessdata.io/specs/data-package/

Chapter 5: Linked open data

'Open first, then link.'

(Caracciolo and Keizer, 2015)

Overview

The process of information discovery and knowledge creation is significantly enhanced by the ability to automatically establish meaningful links between independently produced and managed information resources. This is particularly important in the field of data for development, as the indivisible and inter-linked nature of the SDGs makes it more urgent than ever to join-up a vast amount of information resources and data assets independently owned and managed by many different sectors and communities. The includes, for instance, the ability design data dashboards that can discover relevant links between administrative records maintained by local governments, poverty-reduction performance indicators, and official statistics.

On the other hand, most data providers "are secondary data producers, which means that the journey of a single data point from its origin to its final destination is sometimes not clear to a data user" (Lisowska 2016). By leveraging meaningful, machine-readable links between different datasets over the web, linked data technology can help improve the traceability of data across platforms to their authoritative sources.

This final chapter builds on all aspects of data interoperability that have been explored in previous parts of the Guide. The internet has created unprecedented opportunities for the open exchange of information and the creation of knowledge. The chapter provides guidance on the dissemination of data and statistics on the web according to linked-data principles, including the adoption of the Resource Description Framework (RDF) as a basic data model. It explains the process of creating and maintaining Linked Open Data and provides recommendations and best practices for publishing development data on the semantic web, including the re-use of standard RDF vocabularies and the creation and publication of specialized RDF vocabularies so they can be used by others.

Linked open data on the semantic web

There is a growing interest in tools and technologies that allow for the publication of data in such a way that machines can easily identify and integrate semantically related information resources over the web. Data publishers are increasingly aware of the benefits of applying semantic web technologies to the dissemination of their information assets, so the integration of semantically related data assets can be decentralized to the users and automated with the help of the infrastructure provided by the semantic web.

The linked data approach to data sharing and dissemination refers to a set of best practices for publishing and connecting structured data on the world wide web so it can be automatically discovered and linked together. It consists in marking up information resources (e.g., statistical datasets) with special metadata elements designed to be semantically referenced over the web, thus enabling machines (so called "bots" or "smart data integration agents") to retrieve meaningful information from a global network of "subject-predicate-object statements".

For example, linked-data tools enable users to search the world wide web for official statistics from multiple sources (e.g., different national statistical organizations and international agencies) based on the concept of "gender equality" to return links to authoritative sources of related data and statistics, such as

on the proportion of seats held by women in single or lower houses of parliament, even when the latter is only tagged with the semantically related concept of “empowerment of women”.

A number of metadata elements that have been defined for use on the semantic web can be useful for the development of linked open SDG data. Some pre-defined elements can be used directly while others may need to be semantically mapped (see chapter 3). There is a need for a simple set of metadata elements that can be used to describe SDG data resources (e.g., statistical datasets, maps, data visualizations, etc.) on the web. This metadata has to be useable by non-statisticians and should also help make data resources more visible to search engines on the web.

Linked data principles

The linked data approach to connect datasets across the web was originally proposed by Berners-Lee (2006), putting forward a list of four principles for publishing data online as a collection of machine-understandable statements, ready to be harvested and processed:

1. Use Uniform Resource Identifiers (URIs) as names for things;
2. Use HyperText Transfer Protocol (HTTP) URIs so that people can look up those names;
3. When someone looks up a URI, provide useful information; and
4. Include links to other URIs so people can discover more things.

Exposing data as linked data on the semantic web makes it discoverable by a variety of data communities, enabling multiple users and applications to link them with other information resources and fostering the emergence of a global knowledge network. In effect, linked data transforms the world wide web into a truly global database. These four principles call for making data accessible and related to each other through the web. At heart, they are about promoting interoperability between applications that exchange machine-understandable information over the web (Yu 2011).

Linked data infrastructure

Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a graph-based data model suitable for encoding any kind of data and metadata elements as machine-readable “subject-predicate-object” statements. It provides a general method for describing semantic relationships between data objects, laying the foundation for building and publishing linked data on the web. Expressing data in RDF format allows for effective data integration from multiple sources. Detaching data from its original schema enables multiple schemas to be interlinked and queried as one, and to be modified without changing the data instances. Any data, regardless of their format, can be converted to RDF data.

Web Ontology Language (OWL)

Web Ontology Language (OWL) is an RDF-based markup language designed for publishing and sharing semantic web metadata vocabularies (also known as ontologies). In it, concepts are identified with URIs, labeled with one or more natural languages, documented with various types of notes, and semantically related to each other in informal hierarchies and association networks.

Simple Knowledge Organization Scheme (SKOS)

The Simple Knowledge Organization Scheme (SKOS) is another markup language based on RDF used to express the basic structure and content of thesauri, classification schemes, and similar types of controlled vocabularies. SKOS allows concepts to be published online by referencing them to URIs that can in turn be linked with other data assets and integrated with other concept schemes on the web.

Microdata

Microdata (not to be confused with statistical microdata) is a "specification [that] defines how new HTML attributes embed simple machine-readable data in HTML documents." (W3C 2018) It provides a simple mechanism to annotate existing content with machine-readable labels, using terms from shared vocabularies (e.g., schema.org) so they can be understood by the major search engines on the web. Although Microdata's expressivity is limited, data publishers may consider using it as a stepping stone to start linking data content to commonly available vocabularies on the web.

JavaScript Object Notation for Linking Data (JSON-LD)

JSON-LD is a human-readable format, based on the widely-used JSON format, for the serialization of linked data. It can be easily integrated into existing JSON-based applications. JSON-LD enables the *use* of linked data in web-based programming environments, the development of interoperable REST Web services and applications for the semantic web, and the storage of linked data.

Publishing linked open data

Many organizations are interested in publishing linked open data. However, this is a complex endeavor that requires a gradual approach, especially in situations where resources are scarce and technical know-how and infrastructure need to be developed first. In such contexts, it is recommended that organizations "open data first, and then link" (Caracciolo & Keizer 2015), focusing on priority datasets that are highly visible or which have high reuse value.

In the world of official statistics, it may seem that the semantic web and its related linked data technologies represent a radical disruption to well-established and stable metadata practices. However, the needs that have prompted the development of linked data technologies, including the need to establish common metadata frameworks for data sharing and integration, are the same ones that statisticians have been working to address for decades. Thus, communication and collaboration across disciplines is another critical factor for the successful adoption of linked data principles – in particular, collaboration between experts in statistical, information, library and computer sciences – in order to break down the technical and institutional barriers currently preventing the publication of data and statistics on the semantic web.

Building a roadmap: an interoperability rapid assessment

The following is the first part of the assessment framework produced as part of this Guide. It focuses on the relevance and applicability of conceptual frameworks and the value of institutional frameworks, with a particular focus on legal and regulatory frameworks. It is designed to help inform the development and implementation of data governance strategies and should be supplemented with the resources identified under the Further Reading heading below as well as other context-specific materials.

Action areas	Initial Steps	Advanced Steps
Linking data on the semantic web	Select datasets to be openly linked on the semantic web. Create HTTP URIs to identify datasets. Map the dimensions used to describe the data to existing vocabularies and ontologies.	Where necessary, create HTTP URIs for identifiers of data slices, data points, and individual dimensions, attributes and observation values. Create relationships to broader or narrower terms in other vocabularies.
Publishing open linked data	Publish the original dataset using JSON-LD, Microdata, RDF, or any other format that references the mapped metadata terms. Publish any new concepts using existing vocabularies or ontologies (e.g., SKOS), and make them available on the web. If the new concepts are specializations of existing ones, extend those existing vocabularies with sub-classes and sub-properties derived from base concepts and properties.	Convert and publish the original dataset in linked data formats: RDF, JSON-LD, NT or TTL. Use a triple store to store all RDF data. Serve RDF data over HTTP using the SPARQL Server. Build linked open data applications using semantic web technologies to consume RDF data. Link existing RDF data to publicly available linked data in the cloud (LOD Cloud).
Governance considerations	Announce and promote the use of new vocabularies by registering them with relevant services (e.g., Joinup, LinkedOpenVocabularies, etc.).	Coordinate with other stakeholders working in your domain/field to ensure that there is coherence in the way in which linked open data is being published on the web.

Common pitfalls when implementing open linked data:

- Short-term use cases, aims and objectives tend to cloud the bigger picture. Implementing linked-data approaches requires investment and time, and the benefits may not always be immediately apparent. However, in the longer-term, this approach is likely to increase and enhance the value and usability of development data.
- Ensuring that the meaning of data published through a linked-data approach is not lost or altered can be a challenge. Collaborative approaches to the publication of similar types of datasets across data ecosystems and constituencies are needed to realize the potential of open linked data. Coordination, structure and governance are key.

Further reading on linked open data

- Berners-Lee, T., 2006. *Linked Data - Design Issues*. Available at: https://www.w3.org/DesignIssues/LinkedData.html
- Caracciolo, C. and Keizer, J., 2015. *Implementing Linked Data in Low Resource Conditions*. Available at: http://dublincore.org/resources/training/ASIST_Webinar_20150909/Webinar-Keizer_Caracciolo-2015.pdf

- Coyle, K. (2012). *Linked Data Tools: Connecting on the Web*. Available at: https://www.amazon.com/Linked-Data-Tools-Connecting-Technology/dp/0838958591
- Hedden, H. (2016). *The Accidental Taxonomist*. Medford: New Jersey.
- Lisowska, B. (2016). How can Data Catalog Vocabulary (DCAT) be used to address the needs of databases?, *Joined-up Data Standards Project.* Available at: http://devinit.org/post/can-data-catalog-vocabulary-dcat-used-address-needs-databases/
- Voß, J., 2018. *JSKOS data format for Knowledge Organization Systems.* Available at: http://gbv.github.io/jskos/jskos.html
- W3C, 2014. *The RDF Data Cube Vocabulary*. Available at: https://www.w3.org/TR/vocab-data-cube/
- W3C, 2014. *Best Practices for Publishing Linked Data*. Available at: https://www.w3.org/TR/ld-bp/
- Yu, L., 2011. *A Developer's Guide to the Semantic Web.* Berlin Heidelberg: Springer.

Annexes

Annex A: A Roadmap to Interoperability

Interoperability-friendly Data Governance	Initial Steps	Advanced Steps
Data management, governance and interoperability		
Institutional Frameworks	Identify what model of data governance would work best for your organisation (or you are already a part of) and ensure that interoperability considerations are taken into account from the outset as part of this choice. Put in place a data governance policy that sets out how data is governed across your organisation.	Conduct internal data availability assessments/audits on a regular basis and keep a record of what data is held and handled over its lifecycle. Use this information to periodically review your data governance policy and update it as required. Conduct comprehensive quality assessments and data audits in collaboration with other stakeholders within the local data ecosystem. Develop Monitoring, Evaluation and Learning frameworks that include indicators on data governance issues.
Oversight and accountability	Identify Data Governance Officers (DGOs) and establish Data Stewardship Teams (DSTs) within your organisation.	Convene Data Governance Councils or Data Governance Steering Committees across organizations comprised of technical, operational and support staff, and supported by the Executive to ensure oversight and accountability.
Legal and Regulatory Frameworks (see Annex B for further information)	Identify and map applicable laws and regulations that apply to the data you hold and process. Identify the types of agreements (MOUs, data sharing agreements, service agreements, licenses, etc.) that are best suited to the organization's needs and adopt templates that can be used by staff to share data, procure IT services, etc. Devise corporate policies that incorporate interoperability-friendly approaches and strategies.	Develop bespoke legal templates for contracts, MOUs and licenses that conform to international best practices and are compatible with other frameworks (for e.g. licenses that are compatible with Creative Commons templates). Where resources permit, provide departmental training and sensitization on how to interpret and implement corporate policies.

Common pitfalls in data governance:

- Failing to take an organisational approach to data management and governance issues and relegating 'data' issues to the IT department;
- Not developing/enforcing a clear chain of accountability specifying roles and responsibilities across departments when it comes to the effective governance of data across/between organisations;
- Overlooking/not considering interoperability issues as a requirement when updating or procuring new IT systems; resulting in internal data silos, and multiple types of data held in incompatible formats and schemas;
- Not making the best use of legal and regulatory tools and frameworks that can create a safe and structured environment in which data can be shared and integrated while respecting privacy, data protection and security considerations.

Action areas	Initial Steps	Advanced Steps
	Data and metadata models	
Modelling data structures	Starting from a set of source tables, identify elementary datasets to be modelled (variables or indicators). Identify key entities that are described in the information contained in the dataset (e.g., places, people, businesses…): • Identify the dimensions and attributes needed to describe each entity at the target level of granularity (e.g., location, time period, sex…); • To the extent possible, re-use standard dimensions and naming conventions from existing data models (e.g., from existing SDMX data structure definitions); • Consider merging or splitting columns from original tables to define more useful dimensions for data exchange. Create a separate table of distinct values for each dimension, assigning a unique numeric ID to each row.	Identify or define relevant hierarchies of broader/narrower (parent/child) levels for the dimensions of the dataset. Consider enriching the set of attribute columns with information from other datasets, with a view to making each record self-contained and self-explanatory.
Modelling metadata	Identify a minimum set of metadata elements relevant to describe the dataset.	Map all relevant metadata elements to DCAT vocabulary classes.

Quality considerations	Internally consider user needs and data quality considerations when deciding an approach to modelling.	Establish feedback loops with user groups to enable them to comment on the usability and usefulness of the models that have been employed. Utilise the feedback that is received to regularly update the system and maintain high data quality and usability standards.
Common pitfalls in data modelling: • A common mistake in modelling datasets for sharing and dissemination is to try to replicate internal data structures from operational database systems. The focus should be on producing simple, self-contained datasets that are easy to understand and manipulate by users and client applications.		
Classifications and vocabularies		
Action areas	**Initial Steps**	**Advanced Steps**
Using common classifications and vocabularies	Identify relevant, publicly available, and widely used classifications and vocabularies that can be re-used to codify and populate the content of dimensions, attributes, and measures in a data set. Adopt standard vocabularies and classifications early on, starting at the design phase of any new data collection, processing or dissemination system.	Collaborate with members of other data communities in the development and use of common classifications and vocabularies to describe cross-cutting concepts and relationships. Make any new classifications or vocabularies publicly available and accessible in standard formats, such as CSV, JSON or RDF.
Creating semantic interoperability between classifications	Engage in structural and semantic harmonization efforts, mapping "local" terminology used to designate measures and dimensions to commonly used, standard vocabularies and taxonomies.	Map any custom hierarchies used to group observations at different levels of aggregation to standard classifications.
Governance considerations	Establish policies, methods and procedures to maintain classifications and vocabularies, including rules for adding, changing, moving or deleting terms or relationships, as well as the identification of roles and responsibilities.	Seek the active participation of multiple stakeholder groups in the process of producing and updating standard classification systems and vocabularies.

Common pitfalls when using classifications and vocabularies:

- As information systems become larger, more complex and more interconnected, there is a growing tension between the need to use standard classifications and vocabularies to enhance interoperability, and the need to devise specialized ones for specific user groups. Thought needs to be put into how this balance is set and to the extent possible, efforts should be made to connect new classifications to existing ones to ensure continuity and interoperability down the line.
- Classifications and controlled vocabularies should not be seen as static; they need to be flexible and be continuously updated and adapted to the diverse and changing interests of data producers, data managers, and data users.

Open data formats and standardized interfaces

Action areas	Initial Steps	Advanced Steps
Using open data formats	Make the data available through bulk downloads, for example as CSV files.	Use XML, JSON, GeoJSON or other widely-available open data formats to encode common data elements and sub-elements in such a way that data and metadata are linked together but clearly distinguishable from each other. Use a data containerization format such as the Data Package standard format to publish data sets.
Using standard APIs	Set up a webpage and document all the functionality of existing web APIs in use, describing the resources being provided, the operations that can be performed, as well as the inputs needed for, and outputs provided by, each of operation. Provide additional information such as contact information, any licences used, terms of use, etc.	Develop RESTful web APIs for data dissemination following the OpenAPI specification and document them fully using an API documentation language such as Swagger. Expose datasets using SDMX web services. Serve the description of the API portfolio by a specific end point, to facilitate the discoverability and use of APIs.
Enhancing user experience	Follow common design patterns and rules of communication, so users can easily and intuitively interact with system interfaces.	Create or subscribe to API aggregation services, or “API mashups” to improve developer experience and generate value by eliminating the need to work with multiple APIs.

Common pitfalls when using open data formats and standardized interfaces:

- Over-customization of an interface can inhibit its accessibility, usability and interoperability with other systems. System interfaces should prioritize interoperability and flexibility over specificity and optimization. In short, a balance must be struck between specific user group needs and broader usability.
- Notwithstanding the above, the ability to customize and alter interfaces for specific use cases should be planned for as part of a broader data governance policy.

- All the APIs of an organization should be managed as one product, making sure that all component APIs are standardized and have mutually consistent documentation and functionality, and implement common design patterns built from reusable components.

Open linked data

Action areas	Initial Steps	Advanced Steps
Linking data on the semantic web	Select datasets to be openly linked on the semantic web. Create HTTP URIs to identify datasets. Map the dimensions used to describe the data to existing vocabularies and ontologies.	Where necessary, create HTTP URIs for identifiers of data slices, data points, and individual dimensions, attributes and observation values. Create relationships to broader or narrower terms in other vocabularies.
Publishing open linked data	Publish the original dataset using JSON-LD, Microdata, RDF, or any other format that references the mapped metadata terms. Publish any new concepts using existing vocabularies or ontologies (e.g., SKOS), and make them available on the web. If the new concepts are specializations of existing ones, extend those existing vocabularies with sub-classes and sub-properties derived from base concepts and properties.	Convert and publish the original dataset in linked data formats: RDF, JSON-LD, NT or TTL. Use a triple store to store all RDF data. Serve RDF data over HTTP using the SPARQL Server. Build linked open data applications using semantic web technologies to consume RDF data. Link existing RDF data to publicly available linked data in the cloud (LOD Cloud).
Governance considerations	Announce and promote the use of new vocabularies by registering them with relevant services (e.g., Joinup, LinkedOpenVocabularies, etc.).	Coordinate with other stakeholders working in your domain/field to ensure that there is coherence in the way in which linked open data is being published on the web.

Common pitfalls when implementing open linked data:

- Short-term use cases, aims and objectives cloud the bigger picture. Implementing linked-data approaches requires investment and time, and the benefits may not always be immediately apparent. However, in the longer-term, this approach is likely to increase and enhance the value and usability of development data.
- Ensuring that the meaning of data published through a linked-data approach is not lost or altered can be a challenge. Collaborative approaches to the publication of similar types of datasets across data ecosystems and constituencies are needed to realize the potential of open linked data. Coordination, structure and governance are key.

Annex B: Legal framework definitions, value to interoperability, sources and examples

Class of framework/ agreement	Definition	Value to interoperability	Examples & sources	When to use
Normative frameworks	Broad statements and principles setting a standard of practice or behaviour that *ought* to exist. For example, Principle 1 of the Open Data Charter sets an aspirational standard that data should be 'Open by Default'.	Normative frameworks help set expectations and high-level targets that we should aspire to work towards. There are numerous examples of normative frameworks that apply to the data for stakeholders operating in the development sector which contain provisions and references to interoperability – some explicit others implied.	Fundamental Principles for Official Statistics	To inform the development of interoperability-friendly organisational strategies, and Monitoring, Evaluation and Learning (MEAL) frameworks.
			Open Data Charter	
			Findable, Accessible, Interoperable, Reusable (FAIR) Principles for research	
			Royal Society & British Academy Principles for Data Governance	To create linkages to other organisations and international networks that work on interoperability issues/promote a degree of organisational and data governance convergence.
			African Declaration on Internet Rights and Freedoms	
International laws	Rules that apply to nation states at a global level. International laws can either be set out in international treaties or be the result of custom ('customary international law'), usually long-held and well-established practices. International law exists in different classes, some is binding on states, however in practice many international laws lack enforcement mechanisms making them more akin to good practice recommendations than enforceable laws.	While there are no international laws that explicitly reference 'interoperability', many regional legal frameworks touch upon data protection, data sharing, the right to privacy, data accessibility and other attributes that affect and/or are affected by interoperability. Some regional-specific laws set out quite comprehensive data governance and management regimes, such as within the European Union (EU). The level of binding legal integration across the EU makes it an excellent example of how regional-level interoperable legal	International Covenant on Civil and Political Rights (and two resolutions on the Right to Privacy in the Digital Age)	To inform the development of interoperability-friendly organisational strategies, and Monitoring, Evaluation and Learning (MEAL) frameworks.
			General Data Protection Regulation (EU)	
			Public Sector Information (PSI) Directive (EU)	Where relevant, to ensure compliance with binding international regulations such as the GDPR.
			Council of Europe Convention for the Protection of Individuals with regard to automatic processing of personal data (Convention 108)	
			African Union Convention on Cyber Security and Personal Data Protection	

		regimes can be developed.		
Domestic laws	Binding rules that govern behaviour and relationships within a state. Domestic laws can be national and affect a whole territory or, depending on how a particular state is structured, they can vary between federal states, autonomous regions, developed administrations, etc.	Domestic laws have a key role to play in regulating data-related activities within a country. Effective regulatory regimes can make data sharing and interoperability safer, protective of privacy and ethically acceptable. Conversely, in countries where there are few or no data protection laws, interoperability can be hindered by a lack of regulatory standards.	Digital Economy Act (UK) Data Protection Act 1998 – Particularly Schedule 1 containing the Data Protection Principles Personal Information Protection and Electronic Documents Act (PIPEDA) Canada Model Statistical Law (UNECE, STATSCOM) Federal Trade Commission's (FTC) Fair Information Processing Principles (USA)	With a few exceptions, domestic law will always be binding upon activities taking place within a national territory. In addition to adhering to national laws, it is also important for organisations to understand where *gaps* in the law exist that may impact data sharing and to determine what types of agreements (data sharing contracts, licenses, MOUs, etc.) are best suited to filling those gaps on a case by case basis.
Memoranda of Understanding (MOU)	A form of non-binding agreement that (unlike a contract) can be agreed between more than two parties. In essence, an MOU is a formalized 'promise' that is not legally enforceable.	MOU's can be drafted to contain provisions around how data should be captured, stored, accessed and shared. As such, they can be a very helpful tool for organisations to develop minimally formal agreements with other equal partners to promote interoperability. See for example clause 2.5(i) in the example MOU to the right which specifies the formats that data should be transferred in and the platform on which they should be transferred. MOUs can be particularly helpful in contexts where there are gaps in domestic laws around data protection, sharing, etc. and can help in the short to medium term in creating a regulatory environment that is conducive to	Example Memorandum of Understanding Regarding Data Sharing between the Department for Communities and Local Government (DCLG) and the Department for Energy and Climate Change (DECC) in the U.K. MOU template from www.tools4dev.org	MOUs are often used in intra-governmental settings to define how different MDAs interact or how MDAs interact with some external partners. They are suitable in some instances – for instance between two parts of the same government/large transnational entity or where there are legislative gaps in a developing country setting – but should not be used in lieu of contracts by MDAs to procure services that would be better regulated by legally-binding and enforceable contracts.

		interoperability where one does not yet formally exist.		
Data sharing agreements	Data sharing can be a risky business for many companies and organisations. As a result, specialised 'data sharing agreements' have emerged as a legal way for entities to enter into relationships that define how data will be shared between them. Unlike MOUs, data sharing agreements are legally-binding, meaning that they are enforceable in a court of law.	Data sharing agreements can be used for personal, sensitive and non-personal data so long as they reference any applicable domestic laws. Data sharing agreements are useful contracts for organisations which frequently share data with each other. They are less suited to one-off transfers. Similar to MOUs, but with legally binding force, data sharing agreements can contain interoperability-relevant provisions that relate to anything from how data should be transferred and what security considerations need to be met, to how the data should then be used.	Sample pro forma data sharing agreement	Data sharing agreements should be used when there is an intention for two or more organisations to enter into a legal relationship to share data on a regular/repeated basis.
Licences	A licence is a legal permit to do, own, or use material (including data) created by another person or entity. A licence can be a stand-alone agreement, or it can be granted as part of a broader agreement; for instance, data re-use may be licenced as part of a data sharing agreement or service contract.	Licences are a crucial component of organisational interoperability as they facilitate the accessibility and legal re-usability of data. Licences are a crucial component of 'open' data too and form the legal basis for data re-use; either with or without attribution to the data producer. International best practices recommend the use of open licences that contain	Creative Commons Licenses Open Data Commons Attribution License	Should be used by MDAs, multilaterals, INGOs and other large entities as part of a broader open data policy to facilitate data re-use.
			Open Government Licence (UK)	Although pro forma templates such as the Creative Commons Licenses and Open Data Commons Attribution Licence are good starting points for low resource and low capacity settings, in settings where the resources are available, it is recommended that bespoke licences be produced for data and include provisions that are interoperability-friendly.

		attribution clauses as these clauses help trace data provenance and can enhance trust in the data. Licences themselves can be interoperable. The UK's Open Government Licence for instance was designed to be compatible with Creative Commons Attribution (CC-BY) licences.		Licences themselves should be designed to be compatible with pro forma templates, as is the case with the UK's Open Government Licence.
Corporate policies	Although not technically legal frameworks, corporate policies form part of overarching regulatory frameworks. They set out codes of practice and behaviour that guides how organisations and institutions work, capture business models and outline how they should operate.	When binding on institutional behaviour, corporate policies can have either a positive or a negative impact on interoperability. Guidance can be designed to promote interoperability across whole governments. The UK government for example has devised a set of Open Standards Principles that incorporate 'open source' principles into government digital policy. As a result of endeavors such as this one, the UK's Office for National Statistics has been able to lawfully upload the source code for its SDGs reporting platform to GitHub under an open licence enabling other countries and entities to replicate and adjust the template to meet their own SDG reporting needs. Corporate policies can therefore facilitate	UK open standards principles South Africa's Minimum Interoperability Standards for Government Information Systems (2011)	The development of corporate policies that guide institutional behaviors is an integral part of any organisational process and does not have to be resource intensive. Existing examples of good practice from around the world should be studied as part of the design phase (within a desk review of mapping exercise) of new policies and amendments/insertions then made to incorporate interoperability-friendly practices. Corporate policies also offer opportunities for organisations to incorporate broader good data governance practices into their operations, such as the establishment of Data Governance Councils (mentioned earlier in the Chapter). Where resources permit, departmental training and sensitization on how to interpret and implement corporate policies should be considered and accountability and MEAL structures established.

		interoperability across all four layers: technical, data, human and institutional.		

Bibliography

1. Berners-Lee, T. (2006). *Linked Data - Design Issues*. Available at: https://www.w3.org/DesignIssues/LinkedData.html [Accessed 17 October 2018].
2. Biehl, M. (2016). *RESTful API Design: Best Practices in API Design with REST.* API-University Press.
3. Caracciolo, C. and Keizer, J. (2015). *Implementing Linked Data in Low Resource Conditions.* Available at: https://www.slideshare.net/CIARD_AIMS/implementing-linked-data-in-lowresource-conditions [Accessed: 18 September 2018].
4. DAMA International (2017). *Data Management Body of Knowledge* (2nd Ed). New Jersey: Technics Publications.
5. Feld, C. S. and Stoddard, D. B. (2004). Getting IT Right, *Harvard Business Review*. Available at: https://hbr.org/2004/02/getting-it-right [Accessed 10 October 2018].
6. Goldstein, E., Gasser, U., and Budish, B. (2018). *Data Commons Version 1.0: A Framework to Build Toward AI for Good.* Available at: https://medium.com/berkman-klein-center/data-commons-version-1-0-a-framework-to-build-toward-ai-for-good-73414d7e72be [Accessed 15 October 2018].
7. Halevy, A. and Ordille, J. (2006). 'Data Integration: The Teenage Years', *Artificial Intelligence*, 41(1), pp. 9-16.
8. Halevy, A. et al. (2017). Data Integration: After the Teenage Years, *in* Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. Available at: https://dl.acm.org/citation.cfm?id=3034786 [Accessed 17 October 2018]
9. Hedden, H. (2016). *The Accidental Taxonomist*. Medford: New Jersey.
10. Hoffman, E. and Chamie, M. (1999). *Standard Statistical Classification: Basic Principles.* United Nations: New York.
11. 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].
12. Kwasnik, B. H. (2000). 'The role of classification in knowledge representation and discovery', *School of Information Studies: Faculty Scholarship*, 147. Available at: https://surface.syr.edu/istpub/147 [Accessed 10 October 2018]
13. Lalor, T. (2018). *Generic Statistical Business Process Model,* (Vol. 5), [online]. Available at: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0 [Accessed 13 September 2018].
14. Lisowska, B. (2016). How can Data Catalog Vocabulary (DCAT) be used to address the needs of databases?, *Joined-up Data Standards Project.* Available at: http://devinit.org/post/can-data-catalog-vocabulary-dcat-used-address-needs-databases/ [Accessed 10 October 2018]

15. McKeever, S. and Johnson, D. (2015), 'The role of markup for enabling interoperability in health informatics', *Frontiers in Physiology*, 6, pp.1-10.
16. Nativi, S., Mazzetti, P. and Craglia, M. (2017). 'A view-based model of data-cube to support big earth data systems interoperability', *Big Earth Data* 1(1-2), pp. 1–25.
17. Open Data Watch (2018). *The Data Value Chain: Moving from Production to Impact.* Available at: https://opendatawatch.com/reference/the-data-value-chain-executive-summary/ [Accessed 15 October 2018]
18. Palfrey, J. & Gasser, U. (2012). *Interop: The promise and perils of highly interconnected systems*. New York: Basic Books.
19. Steiner, R. (2014). *Non-invasive data governance: The path of least resistance and greatest success* (1st Ed). New Jersey: Technics Publications.
20. Silverston, L. and Agnew, P. (2009). *The Data Model Resource Book, Volume 3: Universal Patterns for Data Modeling*. New York: Wiley Publishing.
21. Stahl, R. and Staab, P. (2018). *Measuring the Data Universe*. New York: Springer Press.
22. Srinivasan, R. (2017). *Whose global village? Rethinking how technology shapes our world.* NYU Press: New York.
23. Torlone R. (2009). Interoperability in Data Warehouses *in* Liu, L., Özsu, M.T. (Eds) *Encyclopaedia of Database Systems*. Boston: Springer Press.
24. Schematron 2018. Landing page, [online], available at: http://schematron.com [Accessed 17 October 2018].
25. SDMX 2018a. Modelling Statistical Domains in SDMX. Available at: https://sdmx.org/wp-content/uploads/Modelling-statistical-domains-in-SDMX-v2-201806.docx [Accessed 17 October 2018].
26. SDMX 2018b. Guidelines for the Design of Data Structure Definitions. Available at: https://sdmx.org/wp-content/uploads/SDMX_Guidelines_for_DSDs_1.0.pdf [Accessed 17 October 2018].
27. SDMX 2018c. SDMX Content-Oriented Guidelines, [online], available at: https://sdmx.org/?page_id=4345 [Accessed on 17 October 2018]
28. U.S. National Library of Medicine 2018. SNM (SNOMED 1982) – Synopsis. Available at: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNM/ [Accessed 17 October 2018].
29. Vancauwenbergh, S. (2017). 'Governance of Research Information and Classifications, Key Assets to Interoperability of CRIS Systems in Inter-organizational Contexts', *Procedia Computer Science*, 106, pp.335–342.
30. Vasudevan, K. (2017). *Microservices, APIs , and Swagger : How They Fit Together What are Microservices ? Why Microservices?* Available at: https://swagger.io/blog/api-strategy/microservices-apis-and-swagger/ [Accessed: 15 October 2018].
31. Walsh, P. and Pollock, R., (No date). *Data Package Specifications.* Available at: https://frictionlessdata.io/specs/data-package/ [Accessed 17 October 2018].
32. W3C 2014. Data Catalogue Vocabulary. Available at: https://www.w3.org/TR/vocab-dcat/ [Accessed 17 October 2018].
33. W3C 2018. HMTL Microdata. Available at: https://www.w3.org/TR/microdata/ [Accessed 17 October 2018].
34. Yu, L., 2011. *A Developer's Guide to the Semantic Web.* Berlin Heidelberg: Springer.