Maximizing Access to Public Data

Technology + Innovation - March 4, 2019

Striking the Balance Between "Open by Default" and Targeted Data Sharing

This brief is written by Open Data Watch with significant insights from Tom Orrell. It has been produced in partnership with SDSN TReNDS as part of its body of research on public-private data collaboration and sharing.

TABLE OF CONTENTS

A joint project of

TRENDS
Thematic Research Network on Data and Statistics

OPEN DATA WATCH

INTRODUCTION

Background and Context

Sharing information is the foundation of all communication between people, from individual interactions to the relationships between whole societies, countries, and cultures. Whether through language, mathematics, music, visual arts, or–more recently–code, the effect is the same: the shaping of our individual and collective understanding of the world through the exchange of knowledge and experience. The sharing of information and knowledge is becoming ever more critical to the shaping of human experience. The globalization of communications driven by advances in computer processing power, improved internet connectivity and speed, and the almost ubiquitous presence of mobile communications devices are largely responsible for this step change. Huge quantities of diverse forms of information can now be processed and shared at previously unimaginable speeds.

Under the hood of this unfolding digital revolution in the production, sharing, and use of information and knowledge lies an equally important revolution–the data revolution.[i] Data, the building blocks of information, have taken on new significance in recent years, largely as a result of the sheer scale at which they are being produced. In 2018, 2.5 quintillion bytes of data were created every day, with quantities set to increase amid the rolling out of 5G specifications in coming years fueling a whole new generation of internet-connected devices – the so-called Internet of Things (IoT) (Marr 2018). The sharp scale of the change in the production of data globally is captured in the statistic that 90 percent of data in the world today was produced in the last two years (ITU).

The prospective benefits of these trends are recognized in the international development sector. The Data Revolution for Sustainable Development is now well established. (UN Data Revolution Group) New partnerships have emerged seeking to harness the potential of this revolution and a dedicated community of practitioners[ii] operates worldwide exploring how to harness it to achieve development outcomes, such as the Sustainable Development Goals (SDGs). Numerous pieces of research and programs of work now center on the use of data to improve development processes and achieve and monitor the SDGs[iii] and feed into efforts to promote evidence-informed decision-making (Jones 2012). Collectively, the data revolution, SDG agenda, and drive for evidence-informed decision-making have raised the profile of "data for development" generally and the need for more data sharing especially. In turn, the need for more data sharing to contribute to evidence-informed decision-making and the achievement of development outcomes has resulted in experimentation with numerous new models and innovations. Among the examples are the GovLab at New York University's research into the forms of public-private data exchanges (The GovLab 2017), TReNDS' and partners' Contracts for Data

Collaboration (SDSN 2019), and the Open Data Institute's trialing of data trust data-sharing models in the United Kingdom (Open Data Institute 2018).

As quantities of data have increased around the world, calls for publicly-produced data to be made freely available have also increased. New movements and organizations around open data (Open Data Charter), open government (Open Government Partnership), and open knowledge (Open Knowledge International) have emerged over the past two decades to support the public's right to information. This right is further supported by the United Nations Fundamental Principles of Official Statistics, a set of ten principles that lay out the professional and scientific standards for national statistical offices (NSOs). The first principle, which arguably incorporates the remaining nine and embraces the core principle of open data, states:

> "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." — UN Statistics Division 2014

There are also economic reasons for the growth of the open data movement. Government-produced data are public goods and though they may be expensive to produce, they create economic benefits when they are open. Public goods are non-excludable, meaning that their use by one person does not reduce their availability for use by another. As a result, they can be used and reused many times, each time increasing their social and economic benefit from new products and services created or, more indirectly, from efficiency gains and the reduction of transaction costs (Pollock 2010). In one of the earliest studies of the benefits of open data, Rufus Pollock estimated welfare gains to opening data that were previously sold by the British government to range between 1.6 and 6 billion GBP (Pollock 2010). Another study of the European Union's open data portal predicted a total of 1.7 billion euros will be saved in efficiency gains from open data for the public sector in the year 2020 alone (Capgemeni Consulting 2015).

Objectives and scope

This brief has been produced in partnership with SDSN TReNDS, as part of its body of research on public-private data collaboration and sharing. The concept of "data sharing," in this brief is defined as the disclosure of public information at scale as a form of mass data sharing, and should be what administrative authorities aspire to. Despite there being a strong case for sharing, too often the data that governments produce is not shared with the public. The Open Data Inventory (ODIN), an analysis of the availability of indicators that comprise the basic framework of statistical systems, showed that in 2018 only 29 out

of the 178 countries published data in all 21 data categories assessed such as health outcomes, international trade, national accounts, and pollution (Open Data Watch 2017). Furthermore, 55 countries did not publish data in at least five categories, indicating substantial gaps in data sharing.

There are many reasons for failures to share public data. These include: the (sometimes perceived) complexity and effort needed to do so; a lack of political will or incentives to share data; shortfalls in the skills, human and institutional capacity; or a lack of financial resources and investment in public digital and data infrastructures, to name a few. Pervasive and sometimes widening digital divides around the world further stifle efforts to open up and share publicly produced data–or at the very least disincentivize digital data use, and by implication public demand for disclosure–in places where internet access is still prohibitively expensive or digital literacy rates low (World Wide Web Foundation 2016).

This brief adopts a framework (Figure 1) premised on the notion that the value of data can be significantly enhanced through responsible sharing: underpinned by effective and enforceable laws and policies, and cognizant of the need for both sustainable financing for data and investment in institutional capacity and skills.

FIGURE 1: BENEFITS OF OPEN PUBLIC DATA

Sets of principles and norms are emerging to help guide practitioners through this complex and ever-evolving space. One of the key principles that has emerged is that publicly produced data should be "open by default"–the concept that public data should be disclosed unless there is a legitimate reason for it not to be (World Wide Web Foundation 2016). This concept is the cornerstone that sets the stage for how countries and administrative authorities should approach data disclosure and sharing. However, this simple principle hides a far more complex reality.[iv] What should happen in the exceptional cases where data cannot be made open by administrative authorities? What are legitimate reasons for nondisclosure of public information? What are the alternative approaches to data sharing that can maximize public access to data that cannot otherwise be made open?

These are but some of the questions that practitioners working with the data revolution–whether statisticians in NSOs or intergovernmental agencies, development professionals

responsible for knowledge management or open government advocates–are constantly juggling. This brief aims to provide answers to these questions and an overview of the approaches available to authorities seeking to maximize public access to their data. It first explores the applications and limitations of the "open by default" approach to data sharing. It then considers tools available to maximize data sharing when they cannot be made open. It also explores administrative public-to-public data sharing and touches on public-to-private sharing, only insofar as the sharing of data with a private entity relates to the exercise of a public function in situations where public money is being spent. While the converse scenario, private-to-public data sharing, is hugely important to the achievement of development outcomes, it falls outside the scope of this brief but is suggested as an area for further research at the end of the paper, along with other opportunities for future research.

OPEN BY DEFAULT: APPLICATION AND LIMITATIONS

Applying an open by default approach to the disclosure of public data

To understand how an open by default approach to data disclosure can be applied, it is necessary to understand how it is underpinned by access to information (ATI) law, which in some jurisdictions may be referred to as Freedom of Information law, and the legal links between ATI and the concepts of open data and data sharing.

In 1990, only 14 countries had ATI laws (Right2Info 2012). By 2016, the number stood at 112 (Loesche 2017). In most countries with ATI laws, they are the legislative foundation that authorize government bodies to disclose and share information and data with the public, and grant individuals the reciprocal right to access it. ATI laws give effect to the human right to access information, referenced explicitly in Article 19 of the Universal Declaration of Human Rights and recognized under international law as a derivative right of free expression (UNESCO). Many countries' ATI laws include provisions that require governments to "proactively disclose" information,[v] including in the form of open data. In countries with robust ATI laws, the concept of making data open by default then emerges as a preferred policy approach to implementing and operationalizing the legal duty to proactively disclose information and data by creating a presumption in favor of openness–i.e. information and data should be shared with the public unless there is a legitimate reason not to. Over 65 countries have signed up to the Open Data Charter, whose first principle is entitled "Open by Default," thus, committing themselves to this approach.

As the prevalence of ATI laws has expanded, so too have debates about the extent and scope of public duties to disclose information and data. With the disclosure of public

information–typically things such as government budgets, aggregated official statistics, organizational policies, etc.–the ways in which they can be shared with the public are in some respects less complex than those that relate to "data." This is because the term 'information' implies that data have already been structured, analyzed, and interpreted in some way in the process removing any private or sensitive information.

The sharing of raw data via proactive disclosure and through an open by default approach however is another matter. "Open data" is a complicated concept (as Box 1 illustrates) that requires a more nuanced approach to determine where exactly the boundaries of what can and cannot be shared lie. For instance, taking the example of aggregated official statistics mentioned in the paragraph above, while it is important that the public have access to statistical products ("information"), to what degree are they entitled to the underlying data that are used to produce them? Under an open by default approach, does an NSO have a duty to share the microdata that are used to compile official statistics? If so, in what form?

From a development practitioner's point of view, it is clear why this data would be desirable. It would enable analysts to combine multiple datasets using disaggregated characteristics or to look at the distribution of characteristics across a large population in a more precise manner, rather than relying on means and medians.

From a policy perspective, however, it is when the questions above are asked that

BOX 1: WHAT IS OPEN DATA?

In the simplest of terms, 'open data' is data that is licensed for re-use by anyone, free of charge, subject only to discretionary provisions that the source be attributed or that future distribution of the data be sublicensed under a share-alike provision on the same or similar open terms.

There are a number of interrelated technical, legal, policy, and user considerations that contribute to the above definition and facilitate the ability to disclose and share data openly. These considerations are spread out between numerous tools and resources (some of which are listed below) and can be summarized thus:

Technical considerations

· Machine readability: data, and metadata (data about data), should be provided in formats and published to standards that enable them to be processed by a computer.

· Open formats: linked to the above, data should be published in non-proprietary formats that place no legal restrictions on the re-usability of the data (Open Data Charter).

· Interoperability: consideration should be given to the way in which data is classified and defined to ensure that there is consistency in the meaning of data across datasets and that they are comparable (Collaborative on SDG Data Interoperability 2018).

Legal considerations

· Open licensing: the data should be clearly labelled for re-use. The data license should conform to the standards set out in OKI's definition for open licenses (Open Definition 2.1) and should be compatible with other open licenses (Open Knowledge International).

· Attribution: while this is a discretionary criterion, attributing the data to a source or author is helpful if it is anticipated that the data will be reused numerous times as it helps maintain provenance and traceability.

· Share-alike provisions: similarly to attribution clauses, it is good practice to require any further sharing of open data to be sublicensed on the same terms.

Policy and user considerations

· Accessibility: open data should be published online in a way which makes it discoverable and accessible. Data portals are one way providing public access to open data, although increasingly the use of linked-data approaches is recommended where the resources and capacity exist.

· Timeliness and comprehensiveness: Datasets should be published as completely as possible on a regular basis with enough contextual information to enable users to interpret and use the data responsibly.

· User-driven disclosure: for open data to be usable they must be driven by the needs of users. This requires active engagement with those to whom data is being disclosed. This can be achieved through having active feedback mechanisms that enable users to comment on datasets that are released (Orrell 2017).

the limitations of the open by default approach to data sharing becomes more apparent. While some statistical laws and NSO website terms of use will clarify exactly what can and cannot be shared,[xxix] and the Fundamental Principles of Official Statistics make clear that confidential information should never be shared,[xxx] there is still a significant grey area of uncertainty where clearer guidance is needed.

To understand whether it is possible to share data that fall into this grey area, first it is necessary to understand what the legitimate exemptions to the disclosure and sharing of public data are.

The limitations of open by default: Legitimate exemptions

It is important to recognize that there are legitimate exemptions to the open by default approach to data sharing. Public bodies and authorities collect and compile information and data about almost all conceivable dimensions of society, from highly sensitive personal data collected by health authorities to critical intelligence information and strategic data that inform defense policy. While it is perfectly reasonable for states to keep this confidential information hidden from the public, in order for the public to trust that any state-sanctioned secrecy or duty to protect data is conducted in the public interest, these processes must operate as transparently as possible with clear checks and balances in place to prevent abuse. In short, the need for state secrecy or a duty to protect confidentiality in certain situations should not override the principles of transparency and accountability that underpin the notion of "openness," but should operate in tandem with them.

The first step to being transparent and accountable is clearly demarking what classes of information and data are not accessible to the public, explaining why, and ensuring that legally enforceable checks and balances are in place to prevent abuse of the system. Although classes of information that are withheld from the public differ from country to country, internationally recognized standards do exist. For instance, the Organization for Economic Cooperation and Development (OECD) maintains a set of "Privacy Principles" which form the basis of many countries' data protection and privacy laws (OECD 2010). Where possible, exemptions to ATI laws should be subject to a "harm test": i.e. exemptions should exist only where it is foreseeable that disclosure is likely to cause harm in some way, whether to an individual or a vital national interest. They should also be subject to a "public interest override," meaning that a court of law should have the power to override an exemption if it deems that it would be in the public interest to do so on a case-by-case basis.

While there are a number of legitimate exemptions including information and data on national security matters, defense, and international relations, among others, two areas are of particular importance to NSOs, knowledge managers, and other practitioners within the development sector: personal information and confidential commercial information.

Personal and sensitive personal information

Different jurisdictions have distinctive ways of categorizing and handling what can broadly be termed "personal information" In the United States, although there is no single federal law that regulates the collection of personal data, specific federal and state laws refer to personal identifiable information (PII) and sensitive personal information (SPI). In the EU, the General Data Protection Regulation (GPDR) protects personal data and sensitive personal data (EU 2016). PII or personal data generally include data points such as individuals' names, dates of birth, or email addresses. SPI or sensitive personal data include classes of information and data such as medical records, biometric data, and private financial information. Often, independent regulators are appointed to oversee the application of data protection laws to ensure that they enforced appropriately, such as the Information Regulator in South Africa.[viii] Box 2 provides an example of how and when personal and sensitive health data can be shared internally between government departments, and when they cannot.

While ATI laws have flourished around the world over the past 25 years, data protection laws have lagged behind (CNIL). The vast majority of sub-Saharan African and Middle East and North African (MENA) countries in particular lack specific data protection laws, despite the existence of the African Union's Convention on Cyber Security and Personal Data Protection (African Union 2014). This is problematic not only from a rights perspective, but also because it can have a chilling effect on the willingness of foreign entities to engage in data sharing activities in these countries in the absence of a robust regulatory framework. The EU's GDPR, for instance, sets a high bar on data sharing outside of the jurisdiction and requires any third parties handling EU citizens' personal or sensitive personal data to abide by its high data protection standards–a costly and complicated endeavor, but one that places the rights of the individual at its heart (Power 2016). These uncertainties have the potential to stifle and slow cross-border innovation and the application of data-driven technologies to achieve the SDGs where data protection safeguards are lax or nonexistent. The commitment to "leave no one behind" made as part of the 2030 Agenda adds further pressure on the need to clarify and resolve these issues and establish stable regulatory frameworks in which data can be shared safely and responsibly (UN Development Programme 2018).

BOX 2: THE LIMITS OF INTRA-GOVERNMENTAL HEALTH DATA SHARING IN THE UK

In the UK, the Statistics and Registration Service Act 2007 (the Act) grants the Office for National Statistics (ONS) the authority to take a range of decisions around what types of entities it can partner with, how it can source data to compile statistics, and how to release them. The law does this by conferring the ONS with the authority to "do anything which it thinks necessary or expedient for the purpose of, or in connection with, the exercise of its functions" (British Government 2017). Notwithstanding this broad power, the Act contains numerous checks and balances throughout, ensuring that the ONS has the authority to take decisions that relate to the exercise of its legal functions, but that it is also accountable to the legislature and overseen by the executive. In this way, it is an example of a statistical law that balances NSO independence and openness with accountability.

The Act contains numerous provisions that relate specifically to data sharing, including intra-governmental data sharing. Very clear rules about how and when data can be shared between the ONS and the office responsible for civil registration, national health authority, and tax authority are set out. In relation to health data in particular, very strict rules are set out about the classes of data that can be shared by the health authorities with the ONS. Section 43 of the Act sets out these rules in detail, granting permission to the Health Minister to share limited patient registration information with the ONS. The Minister may only share patients' current and previous addresses, date of birth, sex, patient identification number, and history of registration with the ONS. Patients' names cannot be shared. Importantly, the Act states that "the information disclosed [...] may not include any information about the health or condition of, or the care or treatment provided to, any person." Moreover, the information "may only be used [...] for the production of population statistics." A separate entity, Public Health England, is responsible for the production of health-related statistics in England under different sets of– equally strict–rules (Public Health England).

One of the benefits of having very precise and clear rules around intra-governmental data sharing is that they provide the judiciary with a lot of legal certainty when disputes arise. For instance, in 2018 an English legal case between the Migrants' Rights Network, a British civil society group, and the Home Office (the UK's equivalent to an Interior Ministry) centered on the lawfulness of intra-governmental health data sharing (Bowcott 2018). The case revolved around the question of whether a memorandum of understanding between the health authorities and the Home Office authorizing the sharing of patients' health data for the purposes of identifying immigrants being considered for deportation was lawful. The court found that it was unlawful and breached individuals' right to confidentiality. As a result, the government abandoned the policy.

What the above examples demonstrate is that while data disclosure and openness should be the starting point, to maintain public trust in both official statistics and the government's handling of confidential data, it is important to have very clear rules around intra-governmental data sharing, underpinned by enforceable laws. The balance of these attributes is ultimately what creates an open, transparent, and accountable environment in which data can be responsibly and safely shared across government.

Confidential commercial information

In many countries, certain public functions are routinely undertaken by private companies that are subcontracted by administrative authorities. Subcontracted functions can range from infrastructure maintenance (roads, the electricity grid, broadband Internet services, etc.) to the provision of public services such as health and social care, education, or waste management. While strong ATI laws will provide for the disclosure of any information produced through the use of public funds, even where spent by a private entity, they will also often draw a line at the disclosure of information that could compromise private

companies' business models. In the UK for instance, section 43 of the Freedom of Information Act lists commercial interests as a legitimate exemption where the information concerned constitutes a trade secret[ix] or "would, or would be likely to, prejudice the commercial interests of any person[x] (including the public authority holding it)."[xi]

Similarly, ATI laws should ordinarily be aligned with intellectual property legislation and protect copyright belonging to third parties where necessary. The legal interoperability of licensing structures[xii] both within countries and between jurisdictions is therefore of special importance here (see Box 1: What is Open Data?) to ensure that there is consistency in what data are licensed as "open" and what data can justifiably remain closed.

TARGETED APPROACHES TO DATA DISCLOSURE AND SHARING

Now that the application and limitations of an open by default approach to data sharing have been outlined, it is time to return to the question of what options for data sharing exist where there is a grey area between data being open or not shared at all. The remainder of this section covers interrelated practical approaches and tools to data sharing that can be used by NSOs and other entities engaged in development activities to share as much data as possible while respecting the need to protect personal and sensitive personal data as well as commercial confidentiality.

Removing and de-identifying personal and sensitive personal information

There are a number of techniques and tools available to practitioners seeking to make datasets containing personal or sensitive data as open as possible. They range from the fairly crude–severing tables and redacting documents–to the more complex use, notably use of de-identification techniques that can be automated. Microdata–sets of records containing information on individual persons, households or business entities–have potential on their own to fill current data gaps, enable additional disaggregation of populations and localities, establish baselines, or provide ongoing

BOX 3: DE-IDENTIFICATION TECHNIQUES

1. Suppression: achieved by removing a personal or sensitive data field (in essence the same process as redaction but one that can be programmed to be automated for specified fields);

2. Abstraction: the expression of certain data as part of a range; for instance, expressing individuals' ages as part of a range (e.g. 0 to 5 years);

3. Aggregation: achieved through the clustering of data together into larger units–for instance, representing salaries across a range of individuals as an average rather than a series of distinct data points;

4. Pseudonymization and unique identifiers: a variant of anonymization in which data that can be used to identify an individual (name, age, etc.) are replaced consistently with artificially-generated identifiers (so-called unique identifiers);

5. Perturbance: a method of protecting privacy through the changing of certain values while keeping key aggregates constant. For instance, in a dataset where individual salaries are not important but may be used in conjunction with other data in the set to re-identify an individual, it may be possible to perturb the set by randomly changing salary values across the set by equal amounts so that the aggregate remains constant but individual data points become harder to match to other identifying data (O'Hara).

monitoring for sustainable development. Since microdata sets can contain PII, it is important to have a variety of techniques to make them safe to share.

Severability and redaction

A key concept within ATI legislation is the severability, or separability, of datasets. Just because a dataset as a whole may fall within a legitimate exemption–for instance an administrative dataset relating to education enrollment rates that contains personal information such as children's names, ages and sexes–does not mean that parts of the dataset cannot be severed from the main set and shared. It is still possible to sever columns containing personal data from a broader table and share the remainder. Severing datasets can be a useful way of rendering data safe for disclosure, but is not necessarily the most efficient approach to the mass disclosure of information given the time and effort needed to amend each dataset.

Similarly, documents that contain personal or sensitive personal information in text form can be redacted, obscuring or removing sections of text to render them compliant with any duty to protect personal or sensitive information. While redaction is a useful tool, it can be costly and time-consuming, requiring lawyers or trained specialists to trawl through what can be substantial amounts of documentation to remove personal and sensitive data.

De-identification techniques

Although both severability and redaction have useful applications, they also have limitations (as explained above) and are unlikely to be useful approaches for the disclosure and sharing of large quantities of data on a routine basis. De-identification techniques offer a more practical approach that may be more expensive and time consuming to set up initially, but may prove more efficient in the medium- and longer-term given that many can be automated within information systems.

De-identification is the process of removing data and information that can be used to identify individuals from datasets. A sub-set of de-identification includes data anonymization: the manipulation, or changing, of data to remove characteristics that make it harder to identify individuals. Numerous de-identification and anonymization techniques exist. Some key approaches are set out in Box 3.

However, while anonymization and de-identification of datasets is good practice, it is not always enough to keep a dataset private, especially in the case of datasets with a high number of variables. These "high-dimensional datasets"–datasets that have a large number of columns, attributes, and features–can be joined with other datasets to re-

identify participants, as was done by two computer scientists during a Data for Development Challenge. (The GovLab) Extra care should be taken to anonymize and protect these high-dimensional datasets and laws, and new policies should reflect technological risks that can result from re-identifying people.

BUILDING TRUST: BALANCING COMMERCIAL CONFIDENTIALITY WITH OPENNESS

In addition to protecting privacy and sensitive data, the balance between openness and commercial confidentiality must always be struck. The sharing of public data with private entities for the purpose of the performance of a public function must be based on mutual trust: trust on the part of administrative authorities that private corporations will not misuse public data and put individuals or national interests at risk, and trust on the part of private companies that their commercial interests will not be undermined through the disclosure and sharing of any confidential commercial material by administrative authorities. Two types of inter-related data sharing mechanism exist that can help to strengthen trust in a way that still maximizes public access to information and is underpinned by, and guaranteed in, law.

Trusted user frameworks

A trusted user framework can be defined as a system of data access and sharing that grants vetted, trusted users (usually private companies) access to personal or sensitive personal data when certain conditions are met. They can be used by administrative authorities as a way of sharing certain data with private entities for the purposes of performing or contributing to the performance of a public function. For instance, in countries where healthcare provision is split between numerous public and private entities, the use of trusted user frameworks can be beneficial in ensuring that patients' records are transferable and interoperable across health information systems operated by a combination of public and private entities while also protecting its confidentiality.[xiii] Trusted user frameworks are often used in scientific and other research-heavy fields for granting researchers access to otherwise closed data on the condition that they do not disclose any confidential material or data.

Ultimately, trusted user frameworks are about having systems in place that enable trusted users to have access to information on the condition that they commit to only using the data in ways that protect privacy and abide by legal and ethical norms. These frameworks are of particular relevance to data for the SDGs especially in relation to call detail records from mobile phone companies that contain information about call location, length, and other metadata. When anonymized, these metadata can be safely used to understand population movements, spending patterns, and other policy-relevant trends (Naef et al. 2014). Trusted user frameworks are one way through which these kinds of data are being shared, as demonstrated in the Orange Telecom Data for Development Challenge.

In addition to trusted user frameworks, , technology can also be used to keep data safe. The Open Algorithms project (OPAL) solves the problem of data privacy by only "sending the algorithms to the data," so that people given access cannot see the data (which are kept safe by the agency housing them) but can still perform analyses on the data (OPAL Project). This restricted approach, currently being piloted in Senegal and Colombia, allows for the data to be used in a safe manner and serves as a possible model for future efforts.

Data sharing agreements

Data sharing agreements (DSAs) are a class of legally-enforceable contracts that govern how two entities agree to exchange data. Their use can be mandatory in certain jurisdictions under certain circumstances, such as in the EU under the GDPR, and they are widely used in the private sector to establish the scope and parameters for how data should be used. DSAs are especially useful tools in situations where it is envisaged that similar data will be shared between two entities repeatedly on a routine basis.[xiv] Although pro forma templates for DSAs exist,[xv] where possible these agreements should be tailored to the specific needs of a particular situation.

DSAs are a particularly valuable tool to use between public and private entities as they provide a substantial degree of flexibility in allowing the parties to set their contract terms. The Contracts for Data Collaboration initiative between the GovLab at NYU, TReNDS, the University of Washington, and the World Economic Forum was launched specifically to "address the inefficiencies of developing contractual agreements for public-private data collaboration" (The GovLab 2019). It identifies the areas that a data-sharing agreement in the development sector should cover, including among others: the provenance, quality, and purpose of data; security and privacy concerns; roles and responsibilities; and access provisions, use limitations, and governance mechanisms.

It is important to distinguish DSAs from memoranda of understanding (MOUs). MOUs are non-binding agreements, essentially formalized promises, that can be used as the basis of an agreement between two or more entities to share data. However, the fact that they are non-binding means that they cannot be enforced in a court of law, meaning that they do not provide the certainty that DSAs might. They are, however, useful tools to deploy in contexts where data protection laws are weak and few alternatives exist, or for the purposes of intra-governmental data sharing (e.g. between line ministries).

CONCLUSIONS AND AREAS FOR FURTHER RESEARCH

Conclusions

This brief has sought to explore how administrative authorities can maximize access to their publicly held data through a combination of open by default and targeted

approaches. The paper has taken a broad approach to data sharing, arguing that making data open is in itself a form of mass data sharing. While the focus of the brief has been primarily on public-to-public data sharing, it has explored opportunities for public-to-private data sharing where the purpose for which data is being shared relates to the exercise of a public function by a private body, and where public funds are being spent.

Opening data through an open by default policy approach should be the preferred method for administrative authorities seeking to be as open as possible. Notwithstanding this, it is crucial that the limits of this approach are understood and that guidance is provided to practitioners to enable and empower them to take informed decisions about what data should be open, and what should be closed. Moreover, the disclosure of "data"–as opposed to information–is a complex affair given the nuances involved. While the term "information" implies that data has already been structured and organized in some way, data is far more granular and complex.

As a result, even when adopting an open by default approach to disclosure and sharing, the limitations of the approach should be clear and well documented. Any exemptions should be subject to harm tests and public interest overrides to ensure maximum openness and transparency. In the sustainable development sector in particular, the protection of sensitive personal information and commercial confidentiality are particularly relevant exemptions.

In terms of sensitive personal information, this should never be made open except in exceptional circumstances. Numerous de-identification techniques exist and should be routinely employed to protect, remove, or obscure sensitive data. Regarding confidential commercial information, the terms of how data is shared must be underpinned by mutual trust; trust that public information and data will not be misused, or that commercial secrets and trademarks violated. Trusted user frameworks and data sharing agreements are two mechanisms that can be used to safely share commercial information in ways that foster trust and protect confidentiality.

Areas for further research

This brief has sought to lay the groundwork and foundations for how publicly produced data can be safely and responsibly shared to help achieve development outcomes and contribute to evidence-informed decision-making. Notwithstanding the approach adopted here, "data sharing" is an incredibly broad research area and substantial work remains to be done. A number of areas stand out as being relevant to practitioners working in the data for development field and include the following:

1. A serious issue that merits further exploration is what options for responsible public data sharing exist in countries and contexts where data protection, privacy, and access to

information laws are weak or non-existent. What types of alternative mechanisms, if any, can be used to safely and responsibly share data between stakeholders? What is the role of good data management and governance in such situations? Alternatively, in countries where laws exist but they are not explicitly connected, what steps can be taken to align ATI, data protection, and statistical laws to ensure that they operate in tandem to promote openness?

2. Linked to the above point, while this brief has focused on the disclosure and sharing of public data (public-to-public and public-to-private when related to a public function), there is a need for further research exploring the opportunities and risks involved in private-to-public and private-to-private data sharing in the development sector. As data production increases over time and the role of the private sector becomes more urgent to achieve the SDGs and other development outcomes, this need will also become more urgent. What are the incentives that can drive private sector data sharing for public good? How can private sector incentives for sharing be squared with public policy priorities such as the "leave no one behind" agenda and humanitarian principle of doing no harm? What are the specific opportunities and risks involved?

3. Finally, there is a need to step back and take an analytical view of the different innovations currently taking place in this space. The data revolution has given rise to new forms of public-private partnership that have never existed at scale before. Partnerships between NSOs and other administrative authorities, telecommunications and internet service provider companies, geolocation and earth observation specialists, and many others are now flourishing. New types of partnership arguably require new forms and mechanisms to enable responsible and safe data sharing; what do these look like? Further exploration, consideration and analysis of existing innovative approaches–from data collaborative models, to the innovative use of distributed ledger technologies, to data trusts can help inform the future of the field and provide practitioners with examples of both good and bad practice.

ENDNOTES

[i] The digital revolution refers to the shift away from mechanical and analogue technology to digital electronics. A major result of this shift has been an increase in the availability of computing power for communications and processing, and consequently, and a massive increase in the proliferation of data. Sharing is a critical piece of the linkage between these two revolutions as it provides a framework for moving data between users and systems.

[ii] See for instance: http://www.data4sdgs.org

[iii] Some examples include: OECD. (2017) "Development Co-operation Report 2017: Data for Development." OECD. Available at: https://www.oecd-ilibrary.org/development/development-co-operation-report-2017_dcr-2017-en; Sustainable Development Solutions Network. UN. Available at: http://unsdsn.org/wp-content/uploads/2017/09/sdsn-trends-counting-on-the-world-1.pdf; Open Data Watch. (2016) "The State of Development Data Funding." Open Data Watch. Available at: https://opendatawatch.com/the-state-of-development-data-2016/; Gonzalez Morales, Luis. & Orrell, Tom. (2018) "Data

interoperability: A practitioner's guide to joining up data in the development sector." Global Partnership for Sustainable Development Data & UN Statistical Division. 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

[iv] See for instance the blog Is 'Open by Default' too high a bar? published by the Open Data Charter as part of its 2018 review of the Charter's Principles, available at: https://medium.com/@opendatacharter/is-open-by-default-too-high-a-bar-1bc8c0578480

[v] See for example section 4 of India's Right to Information Act 2005, available at: https://rti.gov.in/rti-act.pdf

[vi] For instance, Open Data Watch's Open Data Inventory (ODIN) sets a standard for open terms of use: "Ideally, every data user should be aware of the terms of use governing the dataset they have accessed. Fully open term of use must specify that data can be used, distributed, or modified without change and for commercial and non-commercial purposes with, at most, an obligation to attribute data to the original source." (http://odin.opendatawatch.com/). Many countries also set out lists of official statistics that should be published, such as Mongolia (http://www.en.nso.mn/law/1). Others go further and list the categories of statistics that should be published and the types of data that should be disclosed, including anonymized microdata (see for instance Singapore's Statistical Law: https://sso.agc.gov.sg/Act/SA1973#pr7-).

[vii] Principle 6 makes clear that, "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." Available at: https://unstats.un.org/unsd/dnss/gp/fp-english.pdf

[viii] For example, following the enactment of the Protection of Personal Information Act of 2013 in South Africa, the government established the Information Regulator to "monitor and enforce compliance by public and private bodies with the provisions of the Promotion of Access to Information Act 2000 and the Protection of Personal Information Act 2013." Available at: http://www.justice.gov.za/inforeg/

[ix] At section 43(1) of the Statistics and Registration Service Act 2007

[x] The term 'person' here includes all natural and legal persons, including incorporated private entities

[xi] At section 43(2) of the Statistics and Registration Service Act 2007

[xii] See Annex B of Data interoperability: a practitioner's guide to joining up data in the development sector for instance. Referenced supra at iii

[xiii] See for example: https://www.healthit.gov/sites/default/files/draft-guide.pdf

[xiv] At section 43(2) of the Statistics and Registration Service Act 2007

[xv] For example: https://www.contractstandards.com/public/contracts/data-sharing-agreement

---

REFERENCES

African Union. (2014) African Union Convention on Cyber Security and Personal Data Protection. Available at: https://au.int/sites/default/files/treaties/29560-treaty-0048_-_african_union_convention_on_cyber_security_and_personal_data_protection_e.pdf.

Bowcott, Owen. (2018) Home Office scraps scheme that used NHS data to track migrants. The Guardian. Available at: https://www.theguardian.com/society/2018/nov/12/home-office-scraps-scheme-that-used-nhs-data-to-track-migrants

British Government. (2017) Statistics and Registration Service Act 2007. Available at: http://www.legislation.gov.uk/ukpga/2007/18/contents

Capgemeni Consulting. (2015) Creating Value through Open Data. European Union. Available at: https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.

CNIL. Data protection around the world. Available at: https://www.cnil.fr/en/data-protection-around-the-world.

EU. (2016) The European General Data Protection Regulation. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679.

ITU. ITU towards "IMT for 2020 and beyond." Available at: https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Pages/default.aspx.

Jones, Harry. (2012) Promoting evidence-based decision-making in development agencies. Overseas Development Institute. Available at: https://www.odi.org/publications/6240-promoting-evidence-based-decision-making-development-agencies.

Loesche, Dyfed. (2017) More Countries Adopt Freedom of Information Laws. Available at: https://www.statista.com/chart/11757/more-countries-adopt-freedom-of-information-laws/.

Marr, Bernard. (2018) How much data do we create every day? The mind-blowing stats everyone should read. Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6f98a69660ba.

Naef, Ed, et al. (2014) Using Mobile Data for Development. Cartesian. Available at: https://docs.gatesfoundation.org/Documents/Using%20Mobile%20Data%20for%20Development.pdf.

OECD. (2010) OECD Privacy Principles. OECD. Available at: http://oecdprivacy.org.

O'Hara, Kieron. Transparent Government, Not Transparent Citizens: A Report on Privacy and Transparency for the Cabinet Office. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61280/transparency-and-privacy-review-annex-b.pdf

OPAL Project. Available at: https://www.opalproject.org.

Open Data Watch. (2017) Open Data Inventory 2017. Available at: http://odin.opendatawatch.com.

Open Data Charter. Available at: https://opendatacharter.net.

Open Data Institute. (2018) UK's first 'data trust' pilots to be led by the ODI in partnership with central and local government. Open Data Institute. Available at: https://theodi.org/article/uks-first-data-trust-pilots-to-be-led-by-the-odi-in-partnership-with-central-and-local-government/.

Open Government Partnership. Available at: https://www.opengovpartnership.org.

Open Knowledge International. Available at: https://okfn.org/opendata/.

Open Knowledge International. Open Data Definition 2.1. Available at: http://opendefinition.org/od/2.1/en/.

Orrell, Tom. (2017) What are the principles of joined-up data? Open Data Watch. Available at: https://opendatawatch.com/blog/what-are-the-principles-of-joined-up-data/.

Pollock, Rufus. (2010) Welfare gains from opening up public sector information in the UK. Rufus Pollock. Available at: https://rufuspollock.org/papers/psi_openness_gains.pdf.

Power, Leonie. (2016) Getting to know the GDPR, Part 9 – Data transfer restrictions are here to stay, but so are BCR. Field Fisher. Available at: https://privacylawblog.fieldfisher.com/2016/getting-to-know-the-gdpr-part-9-data-transfer-restrictions-are-here-to-stay-but-so-are-bcr.

Public Health England. Statistics at PHE. British Government. Available at: https://www.gov.uk/government/organisations/public-health-england/about/statistics#our-official-statistics-publications

United Nations Statistics Division. (2014) Principle 1 of the Fundamental Principles of Official Statistics. Available at: https://unstats.un.org/UNSD/dnss/gp/fundprinciples.aspx.

Right2Info. (2012) Access to Information Laws: Overview and Statutory Goals. Available at: https://www.right2info.org/access-to-information-laws.

SDSN TReNDS. (2019) Introducing Contracts for Data Collaboration: New Project on Legal Conditions for Data Sharing. Available at: https://www.sdsntrends.org/blog/2019/1/22/introducing-contracts-data-collaboration.

The Collaborative on SDG Data Interoperability. (2018) A practitioner's guide to joining-up data in the development sector. Available at: http://www.data4sdgs.org/resources/interoperability-practitioners-guide-joining-data-development-sector.

The GovLab. Orange Telecom Data for Development (D4D) Challenge. Available at: http://datacollaboratives.org/cases/orange-telecom-data-for-development-challenge-d4d.html.

The GovLab. (2017) The GovLab at NYU Tandon Launches Website on 'Data Collaboratives – New Forms of Public-Private Data Exchanges that Create Public Value. New York University. Available at: https://engineering.nyu.edu/news/govlab-nyu-tandon-launches-website-data-collaboratives-new-forms-public-private-data-exchanges.

The GovLab. (2019) A repository for strengthening trust, transparency, and accountability: unlocking data for good. New York University. Available at: http://thegovlab.org/new-initiative-contracts-for-data-collaboration/.

UN Development Programme. (2018) What does it mean to leave no one behind? UN. Available at: http://www.undp.org/content/undp/en/home/librarypage/poverty-reduction/what-does-it-mean-to-leave-no-one-behind-.html.

United Nations. About the Sustainable Development Goals. UN. Available at: https://www.un.org/sustainabledevelopment/sustainable-development-goals/.

United Nations Data Revolution Group. Available at: http://www.undatarevolution.org.

UNESCO. Freedom of Information. Available at: http://www.unesco.org/new/en/communication-and-information/freedom-of-expression/freedom-of-information/.

World Wide Web Foundation. (2016) "Closing the digital divide: A briefing note." Available at: https://webfoundation.org/2016/04/closing-the-digital-divide-a-briefing-note/.