Final Report

# Risk to Collaboration
# A Framework to Understand Open Data Sharing Risks

Prepared by Open Data Lab Jakarta of the World Wide Web Foundation

WORLD WIDE WEB FOUNDATION

OPEN DEVELOPMENT INITIATIVE
A PROJECT BY EAST-WEST MANAGEMENT INSTITUTE, INC

SPIDER

This report is financed in whole or in part by Spider. The opinions conveyed are not necessarily shared by Spider. Responsibility for the contents lies exclusively with the author(s) of the report.

# Contents

# Glossary

## Data

A collection of facts in the form of numbers, symbols, and letters, that describe some objects or phenomena.

## Data point

A collection of specific data within the dataset. For example: a data point about “Age” or “Gender” in the population dataset.

## Knowledge Asset

Knowledge assets are data points within a given dataset that are expected to have value, assuming the ability of potential users to leverage that data.

## Risk

Negative events or situations that may jeopardise the success of open data sharing.

# Background[1]

The Open Development Initiative (ODI), a project of the East-West Management Institute (EWMI), "stimulates public demand, builds coalitions, and offers a constantly evolving platform to support the transparent sharing and analysis of data to improve and inform constructive dialogue and decision making for sustainable and equitable development." The project aims to increase awareness regarding key issues in developing countries in this case the Mekong region[2], using open data to support individual analysis and information sharing, and inform rigorous debate in order to achieve sustainable development.

Over the last eight years, the project has worked intensively in the Mekong region, with very tangible results in Cambodia. Open Development Cambodia was launched in 2011 and has since become a primary and trusted source of development data in the country, notably those related to environmental challenges, and more recently those that concern key indicators in the Sustainable Development Goals (SDGs). However, while partnerships have been established in the other four countries, progress variation between countries is largely caused by weaknesses in institutional capacity as well as significant challenges in the enabling environments.

The Swedish Program for ICT in Developing Regions (SPIDER) has supported EWMI in this work. In 2017, SPIDER funded EWMI to implement the project "Open Data to Monitor the SDGs in the Lower Mekong Region" that aims to "investigate the administrative and legal framework surrounding data to track and follow up five SDG goals in the Mekong region and build a pilot framework for how this data can be collected and shared on the Open Development Mekong platform to interested stakeholders".

A key output of this project is the development of a community of open data advocates in Laos, Vietnam, Thailand, Cambodia and Myanmar. EWMI has proposed to build upon its data sharing format by organizing a series of open data workshops with students, academics, local NGOs, and locally-based regional and international groups to build capacity to be able to develop data products consistent with international data standards. This is expected to result in a community of stakeholders with the ability to more actively contribute to OD Mekong and other initiatives locally, regionally, and internationally.

[1] This is adapted from the information that can be found online at https://opendevelopmentcambodia.net/tag/mekong-river/#!/story=post-113868 and https://spidercenter.org/project/open-data-to-monitor-the-sdgs-in-the-lower-mekong-region/

[2] The Mekong region includes the countries of Cambodia, Vietnam, Laos, Thailand, and Myanmar, together with the provinces of Yunnan and Guanxi Zhuang of China. In this case, however, ODI only works in the five lower Mekong countries, excluding China.

## Objective

This study is a continuation of the Phase 1 research, which was conducted to examine the risks and opportunities of open data sharing for ODI's partner civil society organizations (CSOs) in Laos, Myanmar, Thailand, Cambodia, and Vietnam. The current study builds on Phase 1 results and aims to develop an open data sharing risk assessment framework that can assist CSOs in considering the pros and cons of sharing data to the public.

## Method

In this study we develop a framework to assist CSOs in conducting a risk assessment before sharing or publishing data. The development of the framework was made up of two steps. First, a literature review was conducted in order to understand the theoretical approach for building such a framework. Based on the results of the literature review, a risk assessment framework was constructed (see the section on "Proposed framework"). The framework is intended to assist CSOs in determining the data assets contained in the dataset, internal and external benefits to each data user upon possession of the data asset, risk factors and their level of severity, and a mitigation plan to deal with the risks.

Secondly, we tested the applicability of the framework in two main sectors: indigenous data and gender-related data. The experts were asked to test and comment on the usability aspect of the toolkit. In addition, the experts were also asked to think about risk factors specifically relevant to data sharing in the context of indigenous and gender datasets. The findings from each of the research steps is described in the following sections.

## Open Data Sharing Risk Assessment

Data sharing, in the sense of publishing open data, is an increasing trend in the public sector. The main ethical argument for this is that data is produced using public resources, hence it should be used to further the common good. To fulfill the key characteristic of a “public good”, data have to be “non-excludable”, which means that there are no barriers preventing access such that they can be “consumed” by all (Ritchie & Welpton, 2011). Yet, similar to government agencies, CSOs who collect and publish open data are also facing risks. These risks result in the reluctance of CSOs to make their data open.

Similar to risk management methods in other fields, the open data sharing risk assessment also relies on the risk-benefit analysis approach. This approach systematically identifies and manages the risks, while promoting or preserving the benefits that could result from the data sharing activity (CILP, 2016; Eckartz, Hofman, & Van Veenstra, 2014; Zuiderwijk & Janssen, 2015). In this way, the risk assessment process ensures compliance with legal requirements, protection of individual rights and interests, as well as guaranteed data quality.

We have developed a risk-benefit analysis approach for data risk assessment consisting of several critical processes. First, the field of risk management has advocated for the need to understand context prior to identification of the risks (ISO 3100, 2018). Context setting may include understanding the organizational context, strategy, and project goals. Specifically, for an open data risk assessment, prior studies have proposed an identification of the data sharing goal (Eckartz et al., 2014; Zuiderwijk & Janssen, 2015), an assessment of the data inventory including the location where data is stored and identifying individuals with access to the data (Telford & Verhulst, 2016), and an assessment of knowledge assets contained in the dataset. A knowledge asset is a discrete data point contained in the dataset that is expected to have value (Engine Room, 2016; Aljafari & Sarnikar, 2009). As between these elements of context setting, we argue that the assessment of knowledge assets is the most critical element of the identification of risks and benefits associated with the publication of any dataset. The determination of the knowledge assets is critical for the assessor to make calculations on possible risks and benefits of sharing these assets.

Next, the data risk assessment process also involves an identification of risk and benefit factors associated with open data sharing activities. This is because risk management is fundamentally a decision-making process requiring the consideration of emerging risks in light of given benefits. There are many known potential benefits of sharing data. Doing so can maximize the impact of data (or conclusions drawn from it), inform collaboration, provide stronger evidence for advocacy, increase efficiency of service delivery within and outside of an organization, or play a role in decision-making within other organizations, to name a few potential benefits. It is important to understand which of these benefits are relevant in a particular situation. In contrast, the identification and quantification of risk factors is as critical as understanding the benefits but is one of the more complex steps in the risk management process as there is no standard for what constitutes risks to open data sharing.

Finally, the assessment also includes risk mitigation strategies (Eckartz et al., 2014; Telford & Verhulst, 2016). The goal is to identify measures to prevent the materialization of risks.

## Proposed Framework

By building a data sharing risk assessment framework, organizations can more effectively anticipate, prevent and manage emerging "data risks". The next subsections present the proposed framework for an open data sharing risk assessment with its four key components in more detail.

### Step 1: Identification of the Knowledge Assets

Knowledge assets are data points within a given dataset that are expected to have value, assuming the ability of potential users to leverage that data (Aljafari & Sarnikar, 2009). The identification of the list of data assets is a critical first step to identifying the associated benefits and risks.

A knowledge asset can be identified by assessing its (1) value, (2) rareness and (3) imitability. This approach is based on the resource-based view of knowledge and designed to measure the competitive advantage provided by the knowledge asset (Carlsson, 2003).

Does the knowledge of the data enable data users to sense and respond to opportunities and threats in their work? To what extent do other organizations possess similar knowledge assets? And, is the knowledge asset costly and difficult to obtain or imitate for other organizations that do not possess it? These are the key questions to help data owners evaluate knowledge assets.

**Objective:** To identify knowledge asset contained in the dataset.
**Method:**

Identify strategic knowledge assets by assessing their (1) value, (2) rareness and (3) imitability of each data point in the dataset. Below are the key questions to evaluate whether data point can be considered as knowledge asset.

- **Value:** Does the knowledge of the data points enable data users to sense and respond to opportunities and threats in their work?
- **Rareness:** To what extent do other organizations possess similar data point? -
- **Imitability:** Is the data point costly and difficult to acquire for other organizations that do not own it to obtain or imitate?

**Output**: A list of knowledge asset.

### Step 2: Identification of benefits for each stakeholder

Once the data publisher is able to identify the knowledge assets contained in the dataset (Step 1), the next step is to examine the benefits (or incentives) of sharing these data assets systematically and objectively. The benefits need to be evaluated and understood properly at the outset of any risk management process because after mitigating the risks to match the level of risk, the organization (i.e. data publisher) needs to make a decision on whether it is still acceptable to share the data in light of the identified benefits (CILP, 2016).

Open data sharing might bring benefits internally and externally. Sharing data with other CSOs can help data publishers to recognise the gaps in its dataset, leading to a more resilient and trustworthy dataset. Likewise, the external users can use the data for research, development of public policies, and improvement of their problem-solving capacity. More broadly, data sharing could also support the creation of a new public service, stimulating economic growth and innovation (Janssen et al, 2012; Attard, 2015; Canares et al., 2016).

**Objective:** Identification of the benefits upon sharing of the data asset.

**Method:**

Think about the benefit for internal organization and for external users.

The range of benefits should include benefits to:

- **Data publisher** (e.g. Improving data quality, increasing transparency and accountability, increasing outreach and new partnerships)
- **Data user** (e.g. conducting research, developing problem solving ability, supporting secondary sources of data).
- **Society more broadly** (e.g. Increasing citizen participation and knowledge growth, stimulating economic growth and innovation, reducing environmental waste, delivering efficient and effective public services, guarding against terrorism and other crimes).

**Output:** List of potential incentives or benefits for each stakeholder.

### Step 3: Identification and quantification of risk variables

Experts have argued that risk management in the field of data governance has suffered from the absence of any consensus on what constitutes risks of data sharing. Hence, there is currently no single comprehensive list of risk factors of data sharing (CILP, 2016; Zuiderwijk & Janssen, 2015). Nevertheless, prior studies have categorised risk factors using broad categories such as technical versus non-technical risks.

## Risk Identification

Technical risks consist of the risk factors that affect data quality. It is worth noting that in literature, there is no agreement on the dimensions that characterize data quality. Many proposals have been made, but no one has emerged above the others and established itself as a standard (see for example review by Batini, Cappiello, Francalanci, & Maurino (2009)). However, some dimensions are universally considered important, constituting the focus of the majority of the proposals. These dimensions include accuracy, completeness, consistency, and timeliness.

In Redman (1996), accuracy is defined as a measure of the proximity of a data value, v, to some other value, v′, that is considered correct. Completeness is defined as the degree to which a given data collection includes data describing the corresponding set of real-world objects. The consistency dimension refers to the violation of semantic rules defined over a set of data items. An intuitive understanding of such dimensions is suggested by the following examples, referring to a record Citizen, with fields "Name", "Sex" and "Email", as shown in Figure 1. If "Name" has a value "Mke", but "Mike" is the correct value according to a dictionary of English names, this is a case of low accuracy. An example of low completeness is provided by considering "Email". A null value for "Email" may have different meanings, that is (i) the specific citizen has no email address, and therefore the field is inapplicable (this case has no impact on completeness), or (ii) the specific citizen has an email address which has not been stored (in this case the degree of completeness is low). As an example of consistency, let us consider the values of the fields "Name" and "Sex". If "Name" has a value that is "John" and the value of "Sex" is "Female", this may be a case of low consistency. The last dimension, timeliness, is defined as the extent to which data are sufficiently up-to-date for a task.

| **Citizen** | | |
|---|---|---|
| Name | Sex | Email |

**Figure 1:** Example record

Furthermore, with regard to the non-technical factors, there is a strong argument among data protection advocates that data risk management must protect people, and not just data (CILP, 2016). The non-technical factors may be related to political, legal, economic, and social aspects. In contrast to technical risks, non-technical risks may cause harm to individuals, which can be in the form of physical harm, psychosocial/emotional harm, and economic/financial harm (Engine Room, 2016). Physical harms directly place the data publisher as a target and causes physical damage, while psychosocial/emotional harms cause emotional damage to the data publisher and users. Likewise, economic/financial harm causes damage to personal financial assets.

## Risk Quantification

The next step after risk identification is risk quantification using a 3x3 probability-impact matrix as shown in Table 1. For each risk factor identified in step 3, the data publisher will assess the likelihood of occurrence (probability, Pi) and the consequence of risk in terms of what the effect would be if it happens (impact, Gi).

Here, the probability is divided based on percentage of occurrence:

- Low - Assessed as less than or equal to 30% chance of occurrence.
- Medium - Assessed as more than 30%, but less than or equal to 70% chance of occurance.
- High - Assessed as more than 70% chance of occurance.

Following the Data Risk Checker (Engine Room, 2012), impact is divided into Minor, Moderate, and Major according to the extent to which the occurrence of the risk affects (i) data quality recovery cost (costs associated with the re-execution of the process from data collection to publication) due to deterioration of data quality and (ii) severity of harm to data quality and individuals/organizations.

- Minor - Low cost of data quality recovery compared to the original cost of data production with/without direct or indirect threats with minor or low emotional, physical, and/or economic impact. These threats may include verbal aggression, temporary psychosocial distress, temporary economic deprivation, discrediting, or temporary organizational or team breakdown.
- Moderate - Moderate cost of data quality recovery with/without direct or indirect threats with medium to high emotional, physical, and/or economic impact. These threats may include denigration, exclusion of access to civic rights, psychosocial distress, loss of reputation, loss of livelihood, economic deprivation, moderate to severe physical injury with temporary or permanent effects on basic life functions. High impact threats also include organizational infiltration, personal intimidation, persecution, harassment, targeting for rights violations, and organizational or team breakdown.
- Major - High cost of data quality recovery with/without direct threats with catastrophic emotional, physical, and/or economic impact that cannot be mitigated. These threats may include denial of civic rights, detainment, imprisonment, disabling physical injury, or death.

Based on the combination of the probability and impact score, one can obtain the level of criticality for each risk using the equation 1 below (Dani, 2009). The calculation produces three levels of risk criticality: low, medium, high.

Ci = Pi x Gi (1)

| | | **Impact (consequence of risk)** | | |
|---|---|---|---|---|
| | | Minor<br>(1) | Moderate<br>(2) | Major<br>(3) |
| **Probability (Likelihood of occurrence)** | Low<br>(1) | **1** | **2** | **3** |
| | Medium<br>(2) | **2** | **4** | **6** |
| | High<br>(3) | **3** | **6** | **9** |

**Table 1.** Probability-impact matrix.

**Criticality:**

🟩 **LOW (1-2) -** monitor, further analysis (if required)

🟨 **MEDIUM (3-4) -** further analysis required, immediate action.

🟥 **HIGH (>5) –** immediate action, stop.

A criticism of the probability-impact matrix is that although it is relatively easy to use, it is also subjective to bias when assessing risk probability. To mitigate this bias, the tool (Appendix 3) provides a guideline to help the assessor. Another technique to reduce subjective bias is to try different "definition techniques" in describing the scale. The goal is to try different ways of describing the scale to give assessors meaningful frames of reference against which they can estimate the probability of a given risk. At present, a 3x3 matrix is being used, but this can be changed and the scale refined depending on input from users.

**Objective:** Identifying emerging risk factors and quantification of risk rating.

**Method:**

Think through various risks and potential harms that might be inflicted on your organization and staff upon the release of the dataset. Pay attention to the data assets included in the dataset. The risks could include risks associated with either data or people. Consider the list of possible in-country and cross-country data risks (*Appendix 1*).

Evaluate probability and severity of risk using the probability impact matrix.

**Output:** The output of this process is a high-level score (i.e. risk rating) for the organization, with detailed matrices for each type of risk as supporting documentation.

### Step 4: Dealing with risks

Rarely can risks be eliminated entirely. Hence, the last step of the risk assessment strategy deals with prioritisation of the risks presenting the greatest threat to data and people, and identification of measures that can reduce the risk as fully as practicable and prudent in light of the benefits presented by sharing data.

Since it is assumed that the data publisher will not be able to address all the risks at once, it is advisable to prepare a timeline for the mitigation plan (short term, mid-term, and long term). In the case where the risks cannot be mitigated, there has to be a discussion to determine whether to release the data or not.

**Objective:** To identify measures that can reduce or eliminate severity of risk factor.

**Method:**

Prioritise the risks with the highest rating. Among these risks consider what risks can be eliminated entirely. What risks can be mitigated?
How can those risks be mitigated? Please see *Appendix 2* on the resource for risk mitigation.

**Output:** Risk mitigation checklist and timeline

## Testing the Framework

We developed a toolkit that contains all steps of the open data risk assessment proposed in the above section. For step 3, a spreadsheet that automatically maps and colours content according to input was created to assist with the decision-making process.

Several practitioners with experience in conducting open data projects in the Mekong region were invited to test the toolkit between September and October 2019. The results were documented (see Appendix 3). The following table presents the categories of issues identified during the expert review. Four possible challenges for using the toolkit were suggested. In response, we have also proposed solutions to help minimize the challenges.

**Table 2.** Possible challenges when using the toolkit.

| Category | Description | Action taken |
|---|---|---|
| Knowledge about the dataset | This issue is related to the individual's own knowledge of the dataset under assessment. The person conducting the assessment must have the knowledge of the dataset.<br>It is useful for the user to understand what knowledge assets are contained in the dataset, including the benefits and risks of having obtained the dataset. | A knowledge of the dataset is listed as a prerequisite for the assessor. |

| Category | Description | Action taken |
|---|---|---|
| Understanding of the socio-economic and political context of data sharing | This relates to an  understanding of the socio-economic and political context of data sharing within a specific sector, community, country, and region. | An understanding of context, including socio-economic and political aspects, is listed as one of the capacity requirements for the assessor. Also, in the toolkit, examples of possible risks and benefits are given as a guideline for the assessor. |
| Language and technical jargon | There are sections where the language used is quite high-level and technical. | The language has been simplified and the use of technical language has been reduced in the toolkit.<br>In addition, wherever necessary, a definition of the terminology was added in the toolkit to assist the assessor. |
| Technical skill requirement of the assessor | This issue is related to operating the toolkit, which is based on MSExcel. These skills are also needed to fix formatting issues in the toolkit. | Familiarity with MSExcel has been listed as one of the capacity requirements for the assessor. |

The first challenge is related to the user’s knowledge of the dataset. The ability to identify knowledge assets contained in the dataset, envision the benefits of sharing the dataset, and suggest and quantify the risks are all depended on this particular user’s ability. For example, to understand the asset of rarity, the assessor needs to know whether other organizations have similar datasets. To minimize this challenge, it is important to select an assessor with such knowledge, and a second round of review needs to be conducted by a supervisor. The supervisor might also assign other team members to review the results of the assessment, while the supervisor gives the final approval.

Furthermore, data is never neutral. It is influenced by the socio-economic and political factors within the context in which it operates. Hence, the assessor needs to have a wide range of contextual knowledge to envision risks and benefits that may materialize when a particular dataset is shared and used. In the toolkit, examples of possible risks and benefits are given to guide the assessor performing the task at hand. Note that these examples are just a guideline, and the assessor is free to remove the risks that are not relevant to the assessment. They can also add benefits specific to the dataset under review. Overall, while the tool has provided a few examples as a general guideline, the purpose of the open-ended question is to encourage the assessor to think of specific benefits unique to the dataset under review.

The remaining two issues are technical. First, the experts were worried that the language and technical jargon used in the toolkit was too difficult for the target users, especially since the expected users are most likely to have a language other than English as their first language. Some sections of the toolkit used fairly high-level and technical language, and to rectify this, high-level language and technical jargon was reduced in the toolkit. Furthermore, a glossary of terms was added in the report and toolkit documentation to provide definitions for technical terms.

Another technical issue is related to the skills needed to operate the toolkit. Since the toolkit was created on MS Excel, familiarity with the program is needed. For example, skills are needed to resolve formatting issues in the Excel toolkit. As much as possible, we have checked for and removed any formatting and formula issues prior to the publication of the toolkit.

## Conclusion

It is critical that risk management around open data does not continue in the largely *ad hoc,* casual terms that it has evolved into today, although it remains important that the methods remain flexible. Other sectors - for example, finance, civil engineering and environmental management - have seen the development of professional practices of risk management, including specialised research, international and sectoral standards, a common vocabulary, and agreed-upon principles and processes. The same is needed in open data risk management. In some cases, these can be borrowed from areas in which formal risk assessments are better developed, but in others it requires the collaboration of regulators, academics, and civil society to fill important gaps.

This research is an attempt to fill in the gap by proposing a framework that can help open data practitioners link their socio-economic and political contexts in navigating the risks of publishing and sharing their data according to open data principles. While the framework is built on a solid theoretical foundation, the practicality of the toolkit needs more field testing. Hence, it is important to treat both documents as work-in-progress such that further revisions could be made upon the input of users. In the end, improving risk management in open data needs a concerted effort from all stakeholders - from data publisher to users. This report also serves as a call for proposals to come up with new approaches, methods, tools to further improve open data risk assessment, and the development of best practices around the world.

# Reference

Aljafari, R., & Sarnikar, S. (2009). A framework for assessing knowledge sharing risks in interorganizational networks. AMCIS 2009 Proceedings, 572.

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. Government Information Quarterly, 32(4), 399-418.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 16.

Canares, M. P., Marcial, D., & Narca, M. (2016). Enhancing citizen engagement with open government data. The Journal of Community Informatics, 12(2).

CILP. (2016). Protecting Privacy in a World of Big Data: The Role of Risk Management. Retrieved from https://www.informationpolicycentre.com/cipl-white-papers.html

Dani, S. (2009). Predicting and managing supply chain risks. In G. A. Zsidisin & B. Ritchie (Eds.), Supply Chain Risk: A Handbook of Assessment, Management, and Performance (pp. 53-66): Springer.

Eckartz, S. M., Hofman, W. J., & Van Veenstra, A. F. (2014). A decision model for data sharing. Paper presented at the International conference on electronic government.

Engine Room. (2016). Data Risk Checker. Responsible Data Handbook. Retrieved from https://wiki.responsibledata.io/Responsible_Data_Risk_Mapping

ISO 31000. (2018). Risk Management.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258-268.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218.

Ritchie, F., & Welpton, R. (2011). Sharing risks, sharing benefits: data as a public good. Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona, Spain.

Telford, S., & Verhulst, S. G. (2016). A framework for understanding data risk. Retrieved from https://understandrisk.org/a-framework-for-understanding-data-risk/

Zuiderwijk, A., & Janssen, M. (2015). Towards decision support for disclosing data: Closed or open data? Information Polity, 20(2, 3), 103-117.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. COMMUNICATIONS OF THE ACM, 45(4), 211-218.

# Appendix 1: Risk of Open Data Sharing in the Lower Mekong Region

**Table 3.** Risk of In-Country Data Sharing

| Domain (listed in the order of severity (high to low) based on the expert interview) | Risk | Description | Country where the issue is quite prevalent |
|---|---|---|---|
| Legal (Non-Technical) | Copyright violation | Violation of the copyright due to the absence of licensing information on the shared data, sharing data without proper attribution. | Myanmar, Thailand, Vietnam, Cambodia, Lao PDR |
| | Data misuse | Criminal violation because of sharing fake data, altering or misinterpreting data. | Myanmar, Thailand, Vietnam, Cambodia, Lao PDR |
| | Gaps in the regulatory frameworks and requirements | The vague definition of "public" data, gap between regulatory requirements and organisation's preparation to meet them. | Myanmar, Vietnam, Cambodia, Lao PDR |
| Technical | Hacking | Website hacking, third-party access to email communications. | Myanmar, Thailand, Vietnam, Cambodia, Lao PDR |
| | Low-quality dataset | Data available in low quality due to the different formats, duplicated data, and unknown data sources. | Myanmar, Vietnam, Cambodia, LAO PDR |
| | Virus and malware | Virus and malware infected physical media used in data sharing. | Myanmar, Thailand, Vietnam, Cambodia, Lao PDR |
| Political (Non-Technical) | State surveillance | The possibility of being monitored by the authorities. | Myanmar, Vietnam, Lao PDR |
| | Political persecution | Storing and sharing "politically-sensitive' data may be considered against the government. | Myanmar, Thailand, Vietnam, Lao PDR |
| Social (Non-technical) | User's data literacy | Low literacy in collecting, managing, publishing and using data among government officials and the public. | Myanmar, Thailand, Vietnam, Cambodia, Lao PDR |
| | User's online behaviour | Oversharing data/information on social media, failing to use secure communication channels when sharing sensitive information. | Myanmar, Thailand, Vietnam, Cambodia, Lao PDR |

**Table 4.** Risks in Cross-Border Data Sharing

| Domain (Listed in the order of severity (high to low) based on the expert interviews) | Risk | Description |
|---|---|---|
| Legal (Non-Technical) | The absence of legal frameworks | Unavailability of the legal frameworks that can be used for cross-border data sharing. |
| | Differences in legal frameworks among countries | Possible violation due to differences in legal frameworks governing copyright and censorship among countries. |
| Political (Non-Technical) | State monitoring | Authoritarian governments might monitor data communication. |
| | Political prosecution | An individual might be subject to political persecution by sharing information on sensitive domestic issues. |
| Technical | Hacking | Private data breach - communication might be intercepted by third parties, which can lead to identity exploitation. |
| | Unknown data source | Data often comes from an unidentified source resulting in low trust in data validity. |
| Social (Non-Technical) | Limited understanding of the local context | Exposing partners to risk due to limited understanding of their local context. E.g. Sharing data re the Royal family or military junta. |
| | Language barrier | Most laws and regulations concerning data sharing and privacy only exist in the local language. |

# Appendix 2: Resources for Risk Mitigation Strategy

1. Open Data Principles
   a. International Open Data Charter Principles
   b. Eight principles of Open Government Data (OpenGovData.org)
   c. Sunlight Foundation's Open Data Policy Guidelines
2. Data Privacy
   a. General Data Protection Regulation (GDPR) - European Union
3. The Hand-Book of the Modern Development Specialist: Being a Complete Illustrated Guide to Responsible Data Usage, Manners & General Deportment
4. UN Office for the Coordination Humanitarian Affairs (OCHA) "Building data responsibility into humanitarian action"

# Appendix 3: Comments from experts

**Table 5.** Comments from experts review of the toolkit

| Doc No. | Comment | Response | Action (if any) |
|---|---|---|---|
| 1 | "I find it hard to view the dataset and understand the data points, in particular to view the dataset of gender of Cambodia in 2015. All download files have the same name." | This issue is related to the individual's own knowledge of the dataset under assessment. The person conducting the assessment must have the knowledge of the dataset. | Understanding of dataset needs to be listed as one of the capacity requirements for the assessor. |
| 2 | "In the form itself, I am confused sometimes when thinking of possibilities of risks and impacts. Therefore, I am not confident if my responses will be useful or not." | Same as point #1 | Same as point #1 |
| 3 | "I think this toolkit is super interesting and useful, but for the use case scenario described it still seems a bit daunting. It feels more like a tool that would be used at a workshop or a conference to raise awareness, rather than a tool used individually. It definitely seems like a tool that could foster a lot of dialogue, which is ideal and perhaps that is its greater use. At the same time, as a tool, it probably needs to be pushed pretty strongly - I'm envisioning a video on youtube and FB, having it be pinned to the top of our homepage and in the slider, being pinned to the top of CKAN, being pinned to the menu on every page. Is it possible to have this be auto-uploaded onto CKAN with every new dataset?" | Although it is always daunting to conduct a risk assessment, it is important to note that risk assessment is an iterative process. It might need more than one cycle to conduct the assessment, review the results, and making the decision whether to share the dataset.<br>It's also important for the assessor to have sufficient knowledge of the dataset and the context around data sharing. | The toolkit needs to provide a section that lists the capacity requirement for the assessor.<br>A FAQ section will be added in the toolkit guideline to address the reasons why it's important to conduct and also the fact that it should be consultative and the tool iterative over time to meet needs and changing circumstances. |
| 4 | "In addition, the language used is pretty high-level and involves a lot of what seems like data jargon. Is there any way to simplify?" | There are some sections where the language used is quite high level and technical. | To simplify the language and reduce the use of jargon in the toolkit. Ensure that the language is accessible and enables easy translations. |
| 5 | "Readme: formatting issues<br>Appendix 1? Evidence of advice?<br>super detailed - useful but perhaps ambitious? how easily are these ideas translated into local languages? is this meant to be completed individually or as a team? could appear daunting to an | To minimize difficulties of doing the translation, we'll try to reduce the use of high-level languages and technical jargon<br>The risk assessment is conducted individually by a person with | Although the assessment should be done individually with the person who knows the dataset the most and reviewed by a supervisor there could |

| Doc No. | Comment | Response | Action (if any) |
|---|---|---|---|
| | individual with lesser capacity or understanding of how data and risk fit together time estimates for completion might be useful, or maybe a sample? maybe a video of a person completing the assessment?” | knowledge of the dataset under review.<br>The result of the assessment should be reviewed by a supervisor or team (if needed) | be also be provisions for broader consultation should there be some contention that can't be resolved via the toolkit. |
| 6 | “Dataset information:<br>The comments may not be easy to see/understand if the reader is not familiar with excel<br>what are the "data points"? Do you mean the data contained in the dataset?<br>How might a user know about rareness?<br>It might be easier for the questions to be inserted into the assessment instead of "data point 1", "data point 2" etc (I presume that's where the answers go?)” | Yes, data point is the particular data contained in the dataset.<br>To understand the rareness and other key aspect of knowledge asset, the assessor needs to have the knowledge of the dataset including whether other organizations also have a similar data.<br>It is important to note that rareness is only one of the three criteria. A data point can still qualify as a knowledge asset if it meets at least one of the criteria. | Fix the formatting in excel.<br>Add definition of the data point in the tool.<br>Familiarity with MSExcel should be listed as one of the capacity requirements for the assessor. Perhaps add some basic instructions in the readme file. |
| 7 | “Benefit<br>- maybe it would be easier for this to be a checklist, rather than a fill-in box? I know part of this exercise is to get an assessor to consider the dataset more specifically, rather than just checking boxes blindly, but at the same time an empty frame is pretty daunting. This type of exercise seems something that fits better in a capacity building workshop, rather than an everyday tool.” | Although the tool has provided few examples as a general guideline, the purpose of open-ended question is to encourage the assessor to think of specific benefits related to the dataset under review. | Same as the action in point #1.<br>A training could be done to provide an example as target audience of this toolkit won't innately be having capacity to do the assessment. |
| 8 | “Risk<br>I think the equation is not working, it ends up as #Name. Love the form though<br>Under Probability and Impact the dropdown allows choosing a blank - that should be removed<br>Do we also want a choice for "unknown"?<br>There's no colour for the columns in D?” | Thanks for pointing out. | Double check the formula in the excel sheet. |
| 9 | “Readme:<br>This is very useful. Minor point: some texts are hidden I'm not sure if it's because of my Excel. I need to expand specific rows #35, #39 to read the full text. | The excel sheet needs to be reformatted to prevent hidden texts.<br>Appendix 1 are in the report. | Check the formatting of the table |

| Doc No. | Comment | Response | Action (if any) |
|---|---|---|---|
| | I don't see Appendix 1 and Evidence of advice worksheet, are they in different document?" | | |
| 10 | "Dataset Information:<br>On the section: Why the Data Point is considered as knowledge asset? (Provide justification Think in terms of Value, Rareness, and/or Imitability), It might be useful to perhaps create each column for each of the key criteria?" | Some data points might/might not satisfy all criteria. Although, it might be possible to turn the answer into a checklist, the answer might not provide a clear justification to the person who will be reviewing the result (i.e. the supervisor) | A basic tick box response has been added with a follow on comment box. |
| 11 | "Benefit:<br>I think it's important and useful to think about the benefit. I wonder how would this information be used further with this toolkit?" | The information on the benefit will be useful for the management to make the final recommendation. Given the possible benefits, the management will decide whether it is worth to conduct further risk mitigation actions needed before releasing the data. | |
| 12 | "Risk:<br>I like how you make the risk assessment so tangible and concrete. Some language barrier I personally have to read a few time to understand, for example:<br>"Low cost data quality recovery compared to the original cost of data production from collection to publication" and "denigration, exclusion, access to civic rights" Any ways that could make this a little more clear and accessible?" | This has been pointed out as well in point #4. | To fix the language issue. |
| 13 | "I think it might be useful to give some operative definition for "Risk" here or at the beginning in README section. It's great to have broad spectrum of risk. This may not relevant with this toolkit just gonna note here: for people who are not familiar might still unsure and wonder what are "risk" and why it's important to conduct the risk assessment on data sharing. Might be helpful for them to have a brief pointers in README or begging of this sheet." | Thanks for raising the issue. The risk definition will be added to the README sheet | To add risk definition in the readme sheet. |
| 14 | "I have tried filling out the probability and impact, it's great! Seems like the color function doesn't work on my end. It shows | Thanks for raising the issue. | To fix the issue with the Excel formula. |

| Doc No. | Comment | Response | Action (if any) |
|---|---|---|---|
| | #NAME?" | | |
| 15 | "Another minor point is that I'm not sure if legal, political, social factor are non-technical in a broad sense. Perhaps you could categorize them as socio-political factor?" | This is subject to debate. But, we'll leave it as it is for now. | |
| 16 | "Mitigation:<br>It's a really useful exercise to do. Perhaps you could give a few example so people have some guidance." | Appendix 2 will provide guidance for some of the risk factors. | |
| 17 | "Many thanks for sharing the toolkit. I am wondering about sharing data in the region as a whole might be risk in and of itself? In the worksheet titled "Risk", rows NT1 &NT4, read as more state focused than regional. I recall in the risk assessment report, this was flagged as a potential risk. Perhaps a question about the regional impact?" | The risks listed in the table are combination of both in-country and cross-border risks found in Milestone 1. This serves as a guideline. However, the assessor is free to add additional risk items specific to the dataset under review. They can also remove the risks that are not relevant to the assessment | The instruction has been added to the README file and FAQ section. |
| 18 | "For risk factor, it might be the public interest in the data. If people are not interested in using the published data, it will be a waste of time and energy to do." | This could be included as non-technical risk factors. | |
| 19 | " In the 'Dataset Information' Sheet, I would like to add sources of information or data owner in the third column to make sure transparent copyrights.<br>- In the 'Benefit' Sheet, I would like to add Data owner/ provider (especially in case of wisdom/ knowledge of herbal medicine or similar indigenous knowledge)." | Under their current work on responsible data, EWMI has pointed out the importance of having the distinction between data owners and data publisher, although within open data chain, these two actors can be broadly categorized as data supplier | We didn't create a separate row for data owner, but added a note for the user to also specify the benefits for data owner if in case different than data publisher as suggested by the expert.<br>To discuss with EWMI and how to manage this distinction related to their work on Responsible Data Framework. |