User Guide

Open Data Risk Assessment Toolkit

Version 1.0

Table of Content

Revision History

The list below shows the complete revision history of the content.

Revision	Release Date	Description
1.0	October 2019	First draft

Glossary

Accuracy

A measure of the proximity of a data value, v, to some other value, v', that is considered correct. 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.

Completeness

The degree to which a given data collection includes data describing the corresponding set of real-world objects. An example of low completeness is provided by considering data about email address. 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)

Consistency

Refers to the violation of semantic rules defined over a set of data items. If "Name" has a value that is "John" and the value of "Sex" is "Female", this may be a case of low consistency

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.

Timeliness

The extent to which data are sufficiently up-to-date for a task

Introduction

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. 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.

EWMI and The Web Foundation’s Open Data Lab Jakarta developed a framework to assist CSOs in conducting a risk assessment before sharing or publishing data. 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.

Using the toolkit

Below are the step-by-step guideline on how to use open data risk assessment toolkit.

Complete the dataset information sheet

The dataset information sheet contains:

a) Name of the dataset

b) Dataset owner.
   Note: in case that different organization owns the dataset, their approval must be obtained before sharing the dataset
c) Date of assessment
d) Name of the person who completed the assessment
e) Name of the supervisor
   The supervisor is the person who will give the final approval of the assessment.
f) Organization name
   Name of the organization of the assessor and the supervisor

Identify knowledge asset in the dataset

After completing the dataset information, the assessor needs to identify knowledge asset in the dataset. Identification of knowledge assets is a critical step, without which it is not possible to make determination of the possible benefits and risks of sharing the dataset.

Knowledge asset is data point in a given dataset that is expected to have value, assuming the ability of potential users to leverage that data point. Example of data point is “age” or “gender” data within the population dataset. Besides value, other criteria to assess knowledge asset is its *rareness* and *imitability.*

The data point is considered as a knowledge asset if it can satisfy one of the following criteria:

- Value: Does the knowledge of the data point 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?

Hence, in this step, the assessor needs to:

a) Nominate data point(s) in the dataset under review, which may be considered as the knowledge asset.
b) Assess whether each of the data point(s) meets the criterion in terms of *value, rareness,* and *imitability*. Check (√) the box if applies.
c) Provide a brief justification on why the data point meets each of these criteria.

Identify Benefit of Sharing the Dataset

Once the list of data point(s) that can be considered as knowledge asset is known, the assessor can then identify the benefit of sharing the asset. The range of benefit may include benefit to data publisher, for data users, and society more broadly.

a) List the benefit for data publisher (including data owner in case different). This may include improved data quality, achieved greater transparency and accountability, increased outreach and new partnership etc.
b) List the benefit for data user. This may include conducting research, developed problem solving ability, support secondary source of data.
c) List the benefit for society more broadly. E.g. Increased citizen participation and knowledge growth, stimulate economic growth and innovation, reducing environmental waste, delivering efficient and effective public services, guarding against terrorism and other crimes, etc.

How would the information of the benefit be used further by the assessor or the organization? The information on the benefit will be useful for the management to make the final recommendation. Later on, the assessor will also suggest the mitigation activities. These activities are sometimes costly and will also take up organization resources. Based on the knowledge of the possible benefits, the management can decide whether it is worth to proceed with conducting risk mitigation activities before releasing the dataset.

Risk identification and quantification

Risk factors can be classified into broad categories such as technical versus non-technical risks. Technical risks consist of the risk factors that affect data quality. The dimension of data quality includes *accuracy*, *completeness*, *consistency*, and *timeliness* (see the glossary). 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

The following are the steps to identify the risk factor(s) and quantify them:

a) Think through various risks (technical or non-technical) that might be inflicted on your organization and staff upon the release of the dataset. Pay attention to the *knowledge asset* included in the dataset. As a guide, the list of possible risks for open data sharing in lower mekong region is listed in Appendix 1.

b) For each risk factor, the assessor needs to assess the probability of occurrence and impact when the risk is materialized using the

guidelines listed below. To assess this risk you may need to consult a subject expert, legal or custodian.

Guideline for assessing *probability and Impact score:*

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.

c) The toolkit lists the most common open data sharing risks we encounter in the lower mekong region, but the list may not be exhaustive. Another risk (technical or non-technical) may exist that may only apply to your dataset. If this is the case, please add more rows and provide the description of the risk.

d) Conditional formatting has been applied to calculate risk severity rating (Ci) based on the formula below. Three ratings are Minor (Green), Moderate (Yellow), and Major (Red).

$$Ci = Pi \times Gi \qquad (1)$$

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

Criticality:

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

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

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

e) To pass the assessment there must be no red cells in column H. The majority of cells in column H should be green. If the majority of cells in column D are yellow then you should consider whether this dataset requires the Data Programme team to carry out a more detailed risk assessment.

Dealing with risk

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:

a) Prioritise the risks with Moderate rating (Yellow). Note: To pass the assessment there must be no risk with red rating.
b) Among these risks consider what risks can be eliminated entirely. What risk can be mitigated? How can those risks be mitigated? Please see Appendix 2 on the resources for risk mitigation.
c) Prepare a timeline for the mitigation plan (short term, mid-term, and long term).
d) In the case where the risks cannot be mitigated, there has to be a discussion to determine whether to release the data or not.

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

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

Risk of Cross-Border Data Sharing

Domain (Listed in 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

Example of risk mitigation strategy

Category	Strategy	Description	Timeline
Preventive	Only share public data	Data that could be shared should always be included in the category of public data as defined by the law (i.e. Freedom of Information Law)	Immediate term
	Exercise restraint in making public comments	In some countries where state surveillance is rampant, it is advisable to be careful when sharing personal view in social media. In some cases, the authority can trace down the source and even close the open data platform.	Immediate term
	Create regulation for data sharing	Development of a new regulation or improvements in existing laws related to data-sharing (i.e. IT Law, access to public information law)	Medium term

	Seek previous consent	Many of the published data do not have licensing information. In this case, it is important to seek consent from data publisher before sharing the data.	Immediate term
	Improve understanding of the relevant laws	Build awareness on open data, including understanding of relevant laws.	Medium to long term
	Implement data security plans	Although resource constraints and competing demands make it difficult to prioritize and properly manage data security requirements, nonprofits need to take steps to limit vulnerabilities and protecting their organisation from cyber-attacks.[1]	Medium term
	Cross-check data sources	Incomplete information about data sources in the metadata presents greater risks as it reduces the quality and prevents further use of the data.	Immediate term
	Exercise extra care when sharing personal data	Personal data should be handled carefully and the non-profit organisation needs to have policies in collecting and handling personal data.[2]	Immediate term

	Anonymize data before sharing	Data anonymization is needed If publication of some data is prohibited due to privacy protection laws or other reasons that require the publication of anonymized or aggregated data.	Medium term
	Use encryption for data exchange	Encryption scrambles text to make it unreadable by anyone other than those with the encryption keys to decode it. This has become a necessary strategy to use when sharing sensitive information.	Medium term
	Use two-factor authentication	Two-factor authentication is a security mechanism that requires two types of credentials for authentication and is designed to provide an additional layer of validation, minimizing security breaches. Two-factor authentication provides a secondary layer of security that makes it more difficult for hackers to access a person's devices and online accounts.	Immediate term
	Develop cross-border data sharing framework	The policy makers should consider developing a proposal to establish international or regional legal standards for cross-border access and sharing of data. At the same time, the framework needs to promote interoperability in privacy and data protection.	Long term

Detective	Be alert to virus or malicious softwares	It's important for nonprofits to understand the threats of virus and malicious softwares (malwares). They could slow down or cripple systems, and destroy or alter data.	Medium term
	Conduct penetration test for Open Development Mekong platform	Penetration testing is a systematic process of probing for vulnerabilities in IT applications and networks. It is also useful in validating the efficacy of defensive mechanisms and determining how well end-users adhere to security policies.	Immediate term
Corrective	Seek legal advice	The legal expert may provide advice on how to respond to regulatory inquiries, drafting data protection compliance and policies, and defend against enforcement efforts and potential lawsuits.	Ad Hoc. Can be done in an ongoing basis
	Consult a data security expert	Data security experts can assist in helping organisations understand and plan for cyber risks to reduce impact of any future data security incident.	Can be done in an on-going. Related to the penetration test.

[1] https://www.councilofnonprofits.org/tools-resources/cybersecurity-nonprofits

[2] See for example https://the-engine-room.github.io/responsible-data-handbook/

Additional resources:

1. Open Data Principles
    a. International Open Data Charter Principles Eight principles of Open Government Data (OpenGovData.org)
    b. 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"