**ITU AI/ML in 5G Challenge Management Guidelines**

**Part I: Data Sharing Guidelines**

Success of the ITU AI/ML in 5G Challenge depends on the availability of data and whether entities (or data owners) are willing to share data with others. Rapid and unrestricted sharing of data and resources is essential for advancing the Challenge. However, there are cases where unrestricted data sharing is not possible. In this case, this document addresses measures that can be taken to ensure that data providers are able to share relevant data with problem solvers or researchers under specific agreements to ensure data integrity. Therefore, having an institutional data sharing guideline is the first step towards encouraging companies, entities (data providers), collaborators, researchers and professionals to share relevant data for the challenge.

NOTE: Data providers/owners: defined as entities who have data to share for specific problem statements. This data may be useful for training and testing of AI/ML models.

This document outlines a data management and sharing guideline. This guideline would help data owners to derive maximum value from their data while protecting the interests of their institution and its members.

## **Data Classification Categories**

For the purposes of the ITU AI/ML challenge, we consider the data classification categories[[1]](#footnote-1) below:

|  |  |
| --- | --- |
| Data Category | Description |
| Public/Open Data | Data that can be made publicly available because disclosure is associated with little or minimal privacy impact to individuals and/or organizations. This includes data that is anonymous, aggregated and non-sensitive data.  NOTE- This kind of data can be shared without any restrictions. |
| Restricted data | Some data are moderately sensitive, and cannot be shared publicly (as it is) because disclosure can cause minor privacy impact for an individual, put an individual or community at risk of a privacy incident, or negatively impact upon an organization’s capacity to compete in the market or carry out its activities. Example: measurement data obtained per access network or access network site.  NOTE- This kind of data needs to be pre-processed with an intention of removing the privacy impact before being shared.  Restricted data may be available only under a certain conditions set forth by the data provider.  Example-1: Restricted data may be made available after signing a non-disclosure agreement (NDA),  Example-2: Restricted data may be available only for use within the hosted platform and not for moving out of the hosted platform (i.e. no downloading of data may be allowed).  Example-3: Restricted data may be available to citizens of a particular country or region e.g. under data privacy regulations of EU or China. |
| Secret | Also known as “personal, or confidential,” this is composed of highly sensitive information that may cause serious distress or increase risk to an individuals’ safety, or violate an individual’s privacy or impact the compliance to privacy regulations by organizations. This includes personal data that could identify an individual (either on their own or if combined with other data sets), and protection incident management information.  NOTE- This kind of data should be avoided from being shared. |

In order to determine the sensitivity level of a dataset/information type, it is recommended that data owner perform classification of data and risk-assessment on the potential impact that disclosure of each dataset/information type.

For the ITU AI/ML Challenge, we are interested in data that is classified as open or restricted.

## **Options for hosting “restricted data” for AI/ML in 5G Challenge**

Data providers who would like to share data under the “restricted data” category have the following options to choose from;

**Option-1: Self hosted**

* Data providers host ML sandbox, including toolsets (e.g. for training) and data handling. These will be in-premise for data providers.

NOTE - ML Sandbox: defined in [ITU-T Y.3172]

* According to step-7 of “data sharing guideline”, user agreements are drafted for access to this ML sandbox. E.g. No download of data may be allowed.
* According to discussions with participants, a list of interested participants for the problem statement (specific to the data provider) is made by ITU and discussed with the data provider.
* Data provider shortlists the candidates who can access the restricted data.
* User agreement is signed and this makes the participants eligible to compete in the challenge using the restricted data.

**Option-2: ITU hosted**

* Data providers instantiate ML sandbox, including toolsets (e.g. for training) and data handling. These will be in-premise of ITU (Geneva).
* All other steps remain the same as option-1
* NOTE- in this option, ML sandbox maintenance is taken care of by ITU.
* NOTE- ITU hosted ML sandbox may be reused in future editions of such challenges.

NOTE- ITU may facilitate sharing of data between data providers and eligible participants, this may eliminate the need for each participant or team to negotiate the data provider individually.

## **Risk assessment**

Risk assessment must be carried out at an institutional level because data sensitivity is;

* **Contextual**: What may not constitute sensitive data and information in one context, may be sensitive in another.
* **Temporal**: Data may not be sensitive now, but may become sensitive in the future due to changes in context, such as shifts in policies and/or safety of specific populations.
* **Relational**: One dataset on its own may not be sensitive, however it could become sensitive if analyzed in combination with other dataset(s).

## **Classification**

Based on the risk assessment, classification must be carried out at data-set level to identify which of the above data classification categories the data belong to.

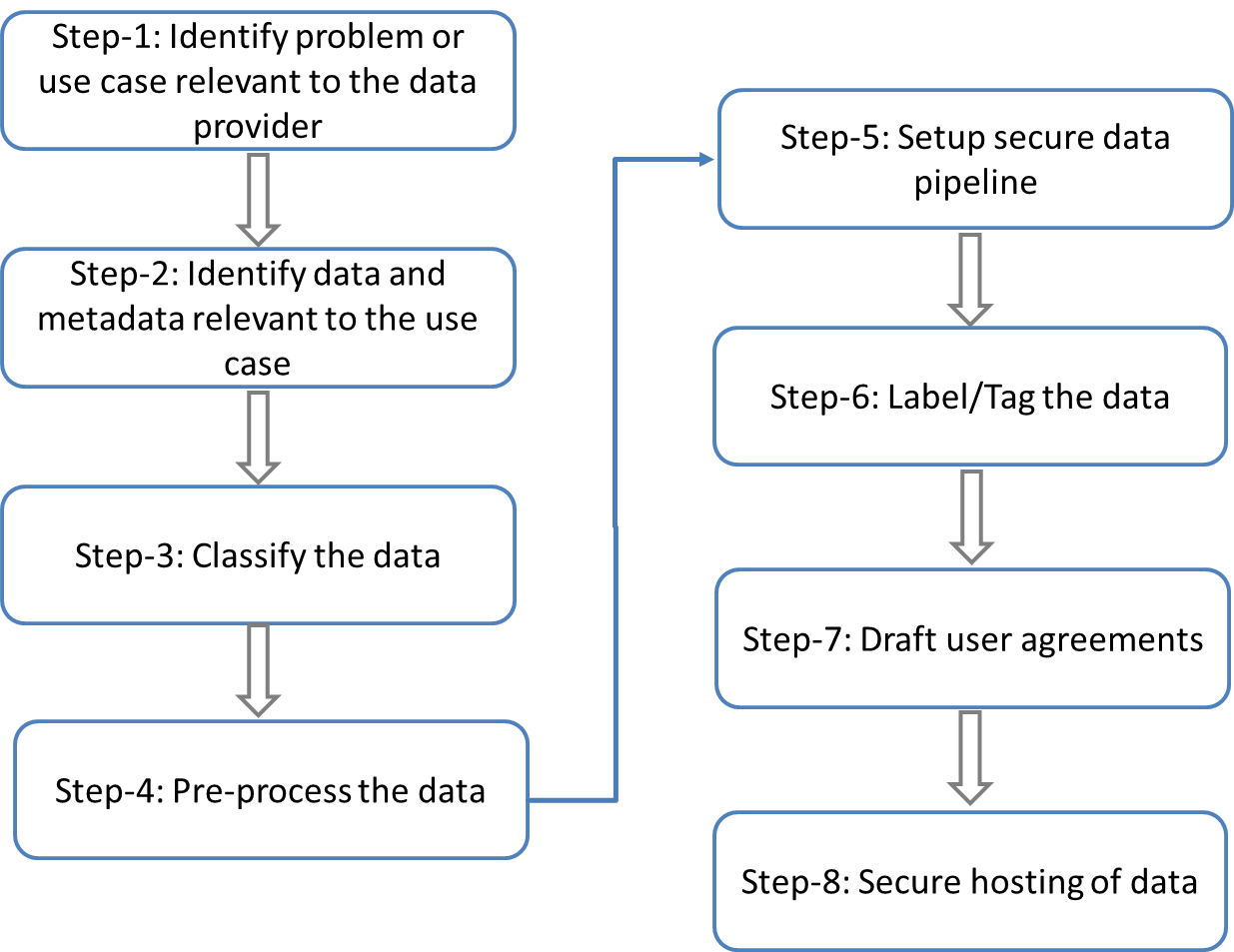
## **Standards, metadata and documentation**

For data sharing to be a success it is important that data are prepared in such a way that those using the dataset have a clear understanding of what the data mean so that they can be used appropriately. To enable this, data owners are encouraged to include with the dataset all the necessary information (metadata) describing the data and their format. This information should include such information as;

* the methodology used to collect data
* definitions of variables
* units of measurement
* data format
* file type of the data
* any assumptions made

## **Data Sharing Guidelines**

The figure below shows the steps to be considered when an entity (data owner) is planning to share data for the ITU AI/ML in 5G Challenge.



Step-1: *Identify problem or use case relevant to the data provider*. In this context, the data owner should choose what type of problem they would like to purse or consider during the challenge. This will help determine the data relevant for the problem.

Step-2: *Identify data and metadata relevant to the use case*. The problem and/or data owner determines what type of data they would provide in order to solve the problem identified in Step-1. In this step, the dataset identified should also contain all the necessary information (metadata) describing the data and their format.

NOTE- ITU can offer expertise to identify data to be collected based on metadata relevant to the use case.

Step-3: *classify the data*. In this step, the data is classified whether it is open (publicly available) or private (provided to challenge participants after certain transformations, under certain rules or user agreements) or secret (not shared at all). This may depend on the internal risk assessment of the data sharing.

Step-4: *preprocess the data.* Thisis an optional step based on the output of step-3 above.Data anonymization is a type of preprocessing whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets. The entity providing data should decide which information to keep for data to be useful and which to anonymize or transform.

Step-5: *Setup secure data pipeline*. A data pipeline is a series of data processing steps. It enables a smooth, automated flow of data from one station to the next. It starts by defining what, where, and how data is collected. It automates the processes involved in extracting, transforming, combining, validating, and loading data for further analysis and visualization. Data pipelines consist of three key elements: a source, a processing step or steps, and a destination. Data pipelines enable the flow of data from an application to a data warehouse, from a data lake to an analytics database, or into a ML pipeline system, for example.

Step-6: *Label/Tag the data. (optional step)* Data labeling is the process of detecting and tagging data samples. The process can be manual but is usually performed or assisted by software. Labeled data is a group of samples that have been tagged with one or more labels. In machine learning, if you have labeled data, that means your data is marked up, or annotated, to show the target, which is the answer you want your machine learning model to predict. In general, data labeling can refer to tasks that include data tagging, annotation, classification, moderation, transcription, or processing. Labeled data highlights data features - or properties, characteristics, or classifications - that can be analyzed for patterns that help predict the target.

Step-7: *Draft user agreements.* A user agreement is an agreement made between the owner, administrator or provider of a service (data owner) and the user of such a service (challenge participants), that defines the rights and responsibilities of both the parties. Privacy policies, terms and conditions, etc. are examples of a user agreement.

Step-8: *Secure hosting of data.* In this step, data owner or ITU provides a platform for challenge to store sensitive data (private or secure data) that in a manner compliant with the entity’s data sharing policy. The challenge participants can access the secure data hosted on the platform by signing non-disclosure agreement or user agreements. This data can be accessed by using password or tokens.

**Part II: Working Model for Judges Panel**

## **Judging Panel**

The Judging Panel is a collection of individuals from across the world who may evaluate, on an ongoing basis, the progress and merit of the solutions proposed by the participants. The Judging Panel will monitor and passively evaluate entries during the Global Round and the Final Conference. The Judging Panel will provide a score for each participant or team at the end of the Global Round. Individuals in the Judging Panel will be selected by the Challenge Management Board.

Working model for the Judges panel

1. CMB will call for and work on the draft of judgment criteria and review it along with the judges panel. This will contain a reference to individual evaluation criteria mentioned in the problem statements and a uniform set of criteria, which when applied, helps to select the top few entries from each regional host. The final reviewed judgment criteria will be baselined in the Challenge website.

NOTE- the final judgment criteria will make sure that the selection of top few entries from each regional host is using a uniform approach world-wide.

1. Judges panel (or its subset) will periodically meet with regional hosts to review the progress of the leaderboard maintained by the regional host. Any best practices or problems which arise out of such meetings will be discussed in the CMB.

NOTE- This will make sure that the selection of global round winners at regional hosts are applying the uniform criteria laid forth by the CMB and judges panel.

1. Judges panel will recommend the invitees from global rounds (hosted by regional hosts) to the final event.
2. Judges panel will review, along with the CMB, the scoring criteria for the entries to the final event. This scoring criteria will be baselined in the Challenge website.
3. Judges panel will recommend the final winners from the event (according to the baselined scoring criteria).

## **Confidentiality terms for judges**

Judges have to accept the terms and conditions laid down by the CMB while joining the judges panel. This includes the following parts:

* By joining the judges panel, the judges accept to abide to confidentiality terms. This means they cannot disclose any information about a submission to any third parties.
* The conflict of interest part:

In case a judge finds conflict of interest in judging some of the submissions, she should declare that well in advance and abstain from judging those submissions. Following scenarios provide guidelines for finding conflict of interest:

Scenario–1: Company or an entity is a member of judges’ panel and an employee of that company submits a solution to one of the problem statements. In this scenario, the judge will declare conflict of interest during the registration or submission. Judge abstains from taking part in the judgment of the submission.

Scenario–2: A person is listed as the contact for a problem statement but she also wants to submit a solution to that problem statement. In this scenario, the host and contact person will declare conflict of interest during the registration or submission. The team is allowed to make a reference solution which will not be considered for any prize for this particular submission, but will be given due consideration of a solution, e.g. an invited presentation can be made by the team based on the submission.

1. IOM, “Guideline for DTM Coordinators: Identifying Sensitive Data and Inter-Organizational Data Sharing Pathways” [↑](#footnote-ref-1)