TERMS OF REFERENCE

Joint Team of AI private company (Developer of Solution) and Academic researcher in the field of Artificial Intelligence for the implementation of an AI Proof-of-Concept

1. Background

Artificial Intelligence (AI) plays a crucial role in fostering comprehensive social and economic development across various sectors. By harnessing the potential of AI technology, we can effectively achieve the objectives of sustainable development and adapt to the challenges of the fourth industrial revolution. Embracing AI enables us to stay abreast of rapid technological advancements and leverage the abundant opportunities it offers to boost economic growth and enhance the performance of governmental entities. AI also creates new job opportunities, contributing to labor market and fosters an environment conducive to innovation and entrepreneurship. Moreover, AI implementation enhances the efficiency, quality, and accessibility of public services while reducing associated costs. These ensure that all segments of society can benefit from improved services and experiences.

The Ministry of Digital Economy and Entrepreneurship (MoDEE) of Jordan has developed and published "AI Strategy and Implementation Plan (2023-2027)" (hereinafter referred to as "the AI Strategy") with the vision of making Jordan a regional leader in the field of AI and providing a unique and attractive technological and entrepreneurial environment for AI to be effective, supportive and an essential component of the national economy.

Under these circumstances, Japan International Cooperation Agency (JICA) has started an international cooperation project with MoDEE titled "The Project for Promoting Artificial Intelligence Ecosystem in the Hashemite Kingdom of Jordan". JICA is dispatching a consultant team to provide technical advice and assistance to the project implementation. The project purpose is to operationalize a **Public-Private-Academia platform** for promoting use of emerging technologies (especially AI) in Jordan. To achieve the project purpose, two major activities are being implemented. One is to establish and improve the capacity of the said platform. Another is to implement PoC (Proof of Concept) programs to promote AI use in real society by means of **Public-Private-Academia consortium**.

The National AI Steering Committee has been established for supervising the implementation of the AI Strategy and Implementation Plan (2023-2027) and takes role to discuss and monitor its progress with selected high-level members from the government (MoDEE), academia, and industry. The JICA project is working closely with the Committee, and the PoC program described in this ToR is selected based on the discussions of the Committee. A Project Secretariat has also been established to help implementation of the project including the PoC program.

This ToR document outlines the requirements for a joint team of private IT company and academic AI researcher to implement a PoC program.

2. Target PoC program

2.1 Title of the PoC program

Data Cleansing of the National Data Warehouse datasets using AI.

2.2 Background and overview of the PoC program

MoDEE has a mandate to gather and store various data from other Ministries at its data center. MoDEE stores those data as a data lake so that the government can perform advanced data analysis and create data dashboard to be used for national strategies and better services to the people in Jordan. MoDEE has dedicated data scientists and data analysts at the data center to perform regular data cleansing and analysis.

However, data provided by other ministries sometimes has rather complex and difficult problems that simple data preparation or cleansing cannot solve. This PoC program is to try to find effective methods for preprocessing and cleansing such datasets. More specifically, this PoC is to research and explore the most appropriate technological procedures (algorithms and/or use of data processing tools) to solve the specific problems in source dataset, so that the processed data becomes ready to be used for any further data analysis, representation (dashboard, etc.), and application. The differences of this PoC from typical data preprocessing / cleansing are as follows.

- Use of deep data processing (possibly by using programming with SQL) – For the first 2 data sets only
- Use of AI in data preprocessing / cleansing / anonymizing

2.3 Important caution of data handling in this PoC

This PoC involves primary data from several government organizations. Since the primary data is the original and extremely important, the primary data must not be altered by this PoC, and the applicant must always work on a copy of original data, and the processed data must not be merged into the original data without approval from both the owner of data (government organization) and MoDEE. This is because any processed data by AI is essentially an "estimation" and cannot guarantee 100% of accuracy. Even if the resulted accuracy of AI process could be very high, there is always a chance to contain false estimated data, and it will be a disaster if those false data will be mixed in to "authoritative" primary data. Therefore, the result of any AI method in this PoC must NOT be mixed with original data.

For the first 2 datasets in this PoC, the applicant must implement two-step approach of Non-AI methods and AI method. In the first step, the applicant must try to apply as many non-AI (deterministic) processes as possible to solve the data problems without sacrificing the 100% accuracy of data (i.e. no estimation should be introduced). For example, the applicant should try to find other relevant data sources (from other government organizations based on the discussion with MoDEE data center) for cross-referencing with the target dataset to determine missing data element in the target dataset (see later for the detail of missing data). As a result of these non-AI processes, the portion of data that has been

recovered or corrected is considered 100% accurate, and it is possible to feedback them to the owner of the original data (government organization). In ideal situation, these non-AI processes may succeed to solve all problems in the data, and there will be no need to apply AI.

However, the first 2 datasets in this PoC are considered to contain problems that cannot be corrected or recovered by non-AI methods only. That's why the applicant should proceed to the 2nd step to apply AI method to solve remaining problems that could not be solved by the non-AI process. See figure below for the explanation of these two-step processes of data preprocessing / cleansing in this PoC.

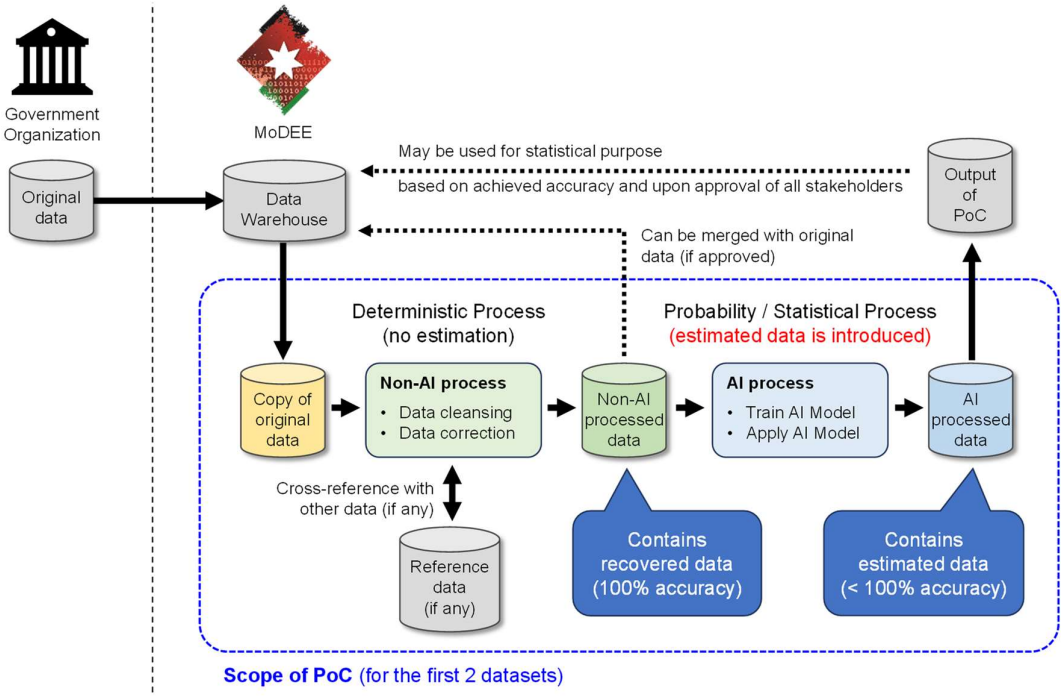

Figure 1: Two steps of data preprocessing / cleansing processes in this PoC (for 2 datasets)

Obviously, no process should be applied to data that has no problem. Non-AI methods should be applied to data that has problems only. As a result of non-AI process, some portions of the data will be recovered or corrected from the data with problems, and there will be remaining data that still has problems. After that, AI processes should be applied to the remaining data (that still has problems) for estimating correct data. See figure below for the target portion of data to apply two types of data cleansing processes.

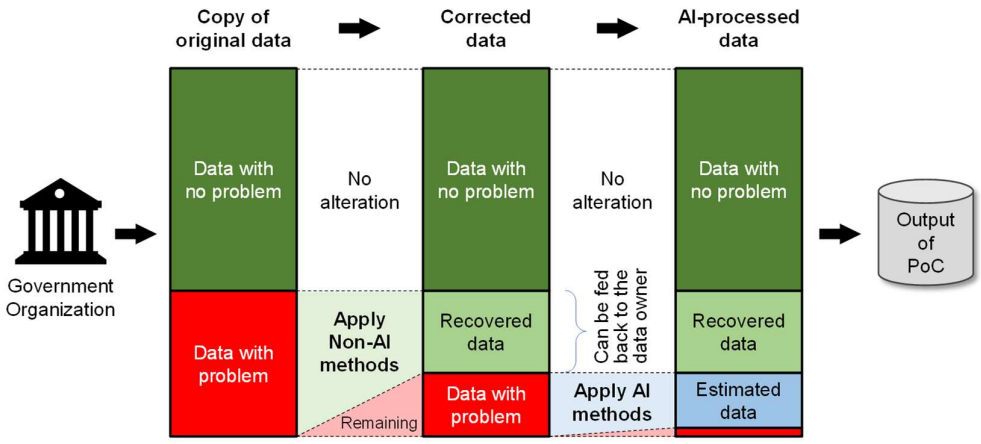

Figure 2: Target portion of data to apply two types of data cleansing processes

Note that the "estimated" portion of data by AI method should NOT be treated as "authoritative" data and should be regarded as a reference only. But the portion can be used for analysis/statistics purposes if its estimation accuracy is high.

2.4 Target use cases (datasets to be preprocessed / cleansed / anonymized) in this PoC

There are 3 separate use cases (datasets to be preprocessed / cleansed / anonymized) in this PoC as summarized in the table below. All these 3 cases are mutually independent, and each of them has specific data problems to be solved by the PoC.

Table 1: Summary of target use cases (datasets) of the PoC

No.	Use Case	Organization in charge	Problem of the data	Non-AI methods	AI methods	Expected goal
1	Determining national ID number in Tawjihi Data (1985-2004)	Department of Examinations and Tests, Ministry of Education	Tawjihi Data (pre-2005) does not contain national ID numbers, and therefore cannot be linked to the Sanad application.	Required (SQL operations, stc.)	Required	Find national ID numbers of all students in Tawjihi data (pre-2005)
2	Linking municipality land data to DLS data	Department of Lands and Survey (DLS)	Data at municipalities does not match well with DLS data	Required (SQL operations, etc.)	Optional	Correctly link municipalities data to DLS data via DLS key
3	Anonymi-zation of court decisions	Ministry of Justice	With the Personal Data Protection Law, personal information in court decisions must be anonymized.	Optional (Standard data cleansing, etc.)	Required	Develop an AI prototype to automatically anonymize court decision while preserving the legal meaning and context

Use case 1: Determining national ID number in Tawjihi Data (1985-2004)

(1) Background

The public secondary data (Tawijhi data) at Department of Examinations and Tests, Ministry of Education is very important as it is in high demand in the services of issuance of graduation certificates and high school transcript. Its data format and contents have been changed several times in the history as follows.

Table 2: History of the format and content of Tawijhi data

Year	Format	Content (Columns)	# of records	Note
2005- *1	Electronic	name, national ID *2, nationality, date of birth, branch, overall mark *3, etc.	~ 5 million	Using legacy system
1985-2004	Electronic	name, year of birth (1985-1999) or date of birth (2000-2004), place of birth, etc. but lacks national ID	1,885,000	Many issues in data quality
1962-1984	Paper	Paper certificate and Paper high school transcript only		(out of scope of this PoC)
1951-1961	Paper	Same as above but in older format (for secondary school)		(out of scope of this PoC)

*1 … This data is uploaded to the cloud and updated twice a year by MoE, and stored at MoDEE data center.

*2 … Data of non-Jordanians born in Jordan or children of Jordanian mothers is linked to a serial number (not a National ID number), while data of non-Jordanians not born in Jordan is not linked to any number.

*3 … Details of exam marks (required for high school transcripts) are stored in the legacy system of MoE . It has been proposed to start uploading high school transcripts (for electronic transcript issuance service).

The main target of this PoC is the dataset of 1985-2004 while the dataset of post-2005 could also be used for training purpose of AI model in this PoC (if required), and then the trained AI model can be applied to 1985-2004 dataset.

(2) Problems of the 1985-2004 Tawjihi dataset

- No national ID number (This is the biggest problem to be solved in this PoC).
- Only birth year is available (no birth month or birth date).
- Inconsistent writing of birthplace, where the same region may be written in multiple ways (for example, “the capital” or “Amman”).
- Inconsistent writing of student names, where names may be written in 3, 4, or 5 parts in the case of different nationalities, and may contain some spelling errors, such as writing the same name in multiple ways like (“Roba” / “Ruba”).
- Some individuals may have updated or changed their family name without updating the data in the Department of Examinations and Tests database.

(3) Goal of Use case 1

- **Goal 1**: Determine as many national ID numbers in 1985-2004 Tawjihi dataset as possible

by cross-referencing the data with CSPD (Civil Status and Passports Department) data using deterministic non-AI methods. Possible non-AI methods include categorizing the data by specific columns and narrowing down the number of candidate records to be matched. The applicant should propose the strategy to perform these deterministic methods in their proposals. Other related data (such as higher education graduates data) may also be requested from other organizations[1] to help identifying the national ID numbers.

- **Goal 2**: Apply AI method to estimate matching CSPD records for 1985-2004 Tawjihi dataset by training the AI model with post-2005 dataset and apply the trained AI model to 1985-2004 Tawjihi dataset. Compare the result with non-AI methods and evaluate the performance of AI model (estimation accuracy, etc.).

Note: Determining national ID is the only goal in this use case. Other problems listed above can be automatically solved once the national ID has been determined.

(4) Procedures to achieve goals

For Goal 1:

1. Examine the 1985-2004 Tawjihi dataset and confirm / analyze the problems in the dataset.
2. Cleans the dataset based on the analysis of step 1. Note that no "data estimation" or "data augmentation" should be introduced at this step unless otherwise discussed and agreed with MoDEE and MOE.
3. Elaborate possible non-AI data matching methods (algorithms) to be applied to the dataset. All methods must be deterministic (no estimation) and must logically ensure 100% accuracy of the matched records. Typically, a non-AI method might be a computer algorithm (script) to be applied to the dataset by using data manipulations such as SQL.
4. It is expected that single non-AI method would not recover all data (depending on the algorithms), so multiple methods should be applied subsequently to remaining data portion that could not be recovered by the previous method through "trial and error" manner as shown in the flowchart below.

[1] Availability of data from other organizations is subject to discussion and agreement among all parties.

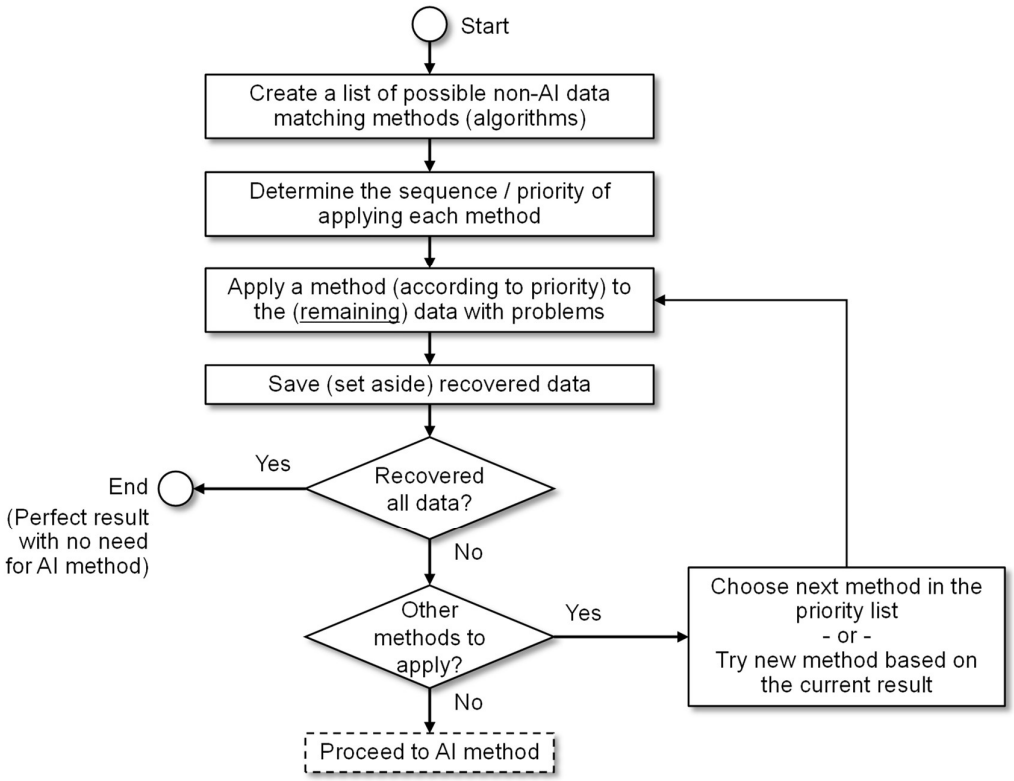

Figure 3: Flowchart of applying multiple non-AI methods to the dataset

(Example) MoDEE data center has already experimented a matching method to a subset of Tawjihi dataset (2002-2004: 330K records) as follows.

- Match the dataset with CSPD by the combination of "birth year" and "unique name".
  - Step 1: Extract CSPD records that has the same birth year as 2002-2004 Tawjihi dataset.
    → 1.1M records were extracted from CSPD data.
  - Step 2: Extract CSPD records that has unique name (unique combination of name parts)
    → 1.0M records were extracted from CSPD data.
  - Step 4: Match the CSPD records of unique names with 2002-2004 Tawjihi dataset.
    → 82% of Tawjihi dataset matched with CSPD data (i.e. National ID were determined).

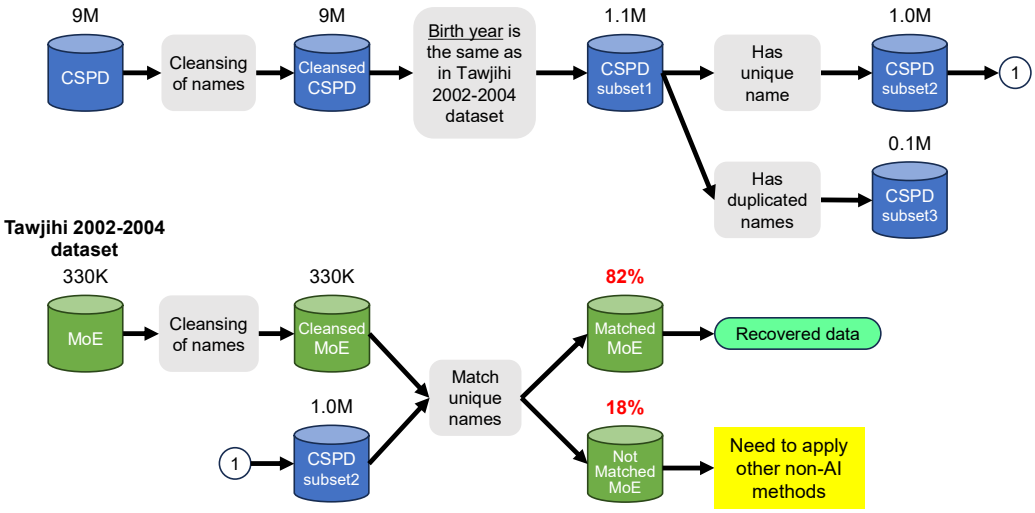

Figure 4: Example of applying non-AI data matching method at MoDEE and its result

Note 1: As you can see in this example, single non-AI data matching method cannot achieve near-100% recovery of the data, and you should apply different methods to "remaining" data that you could not determine National ID numbers.

Note 2: Subsequent non-AI methods should be better selected based on the result of previous method. The choice of method (or algorithm) should be done dynamically by observing the performance of applied methods (in trial-and-error manner) because the "remaining" data is already a "filtered" data by previous methods and thus the possibly effective next method depends on the previous methods as shown in the figure below.

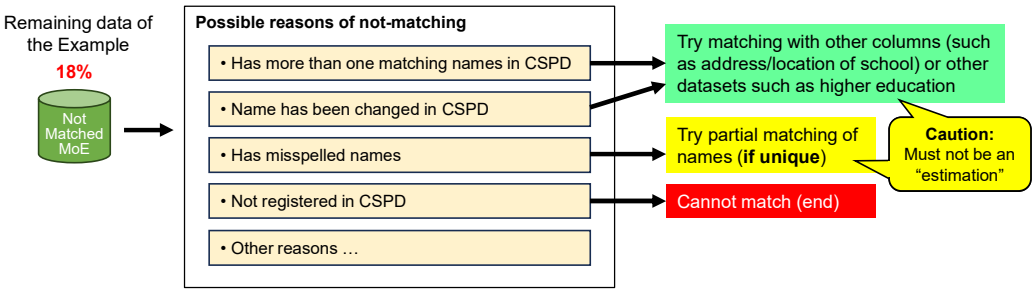

Figure 5: Choosing next non-AI method based on the result of previous methods (example)

For Goal 2:

1. From the post-2005 Tawjihi dataset, create a training dataset for AI model by deriving a dataset that has identical data format as 1985-2004 Tawjihi dataset (i.e. no National ID number, birth year only, etc.) and their corresponding CSPD records as shown below.

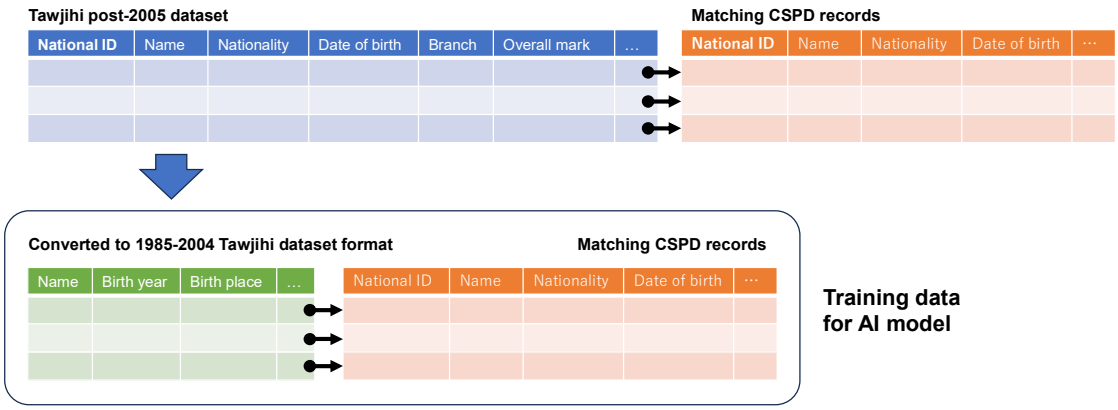

Figure 6: Creating a training dataset for AI model

2. Train an AI model with the created training dataset.

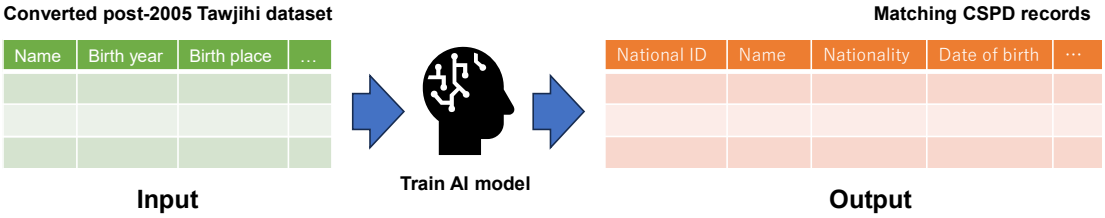

Figure 7: Train AI model using converted post-2005 dataset

3. Evaluate performance of the trained AI model with 1985-2004 Tawjihi dataset

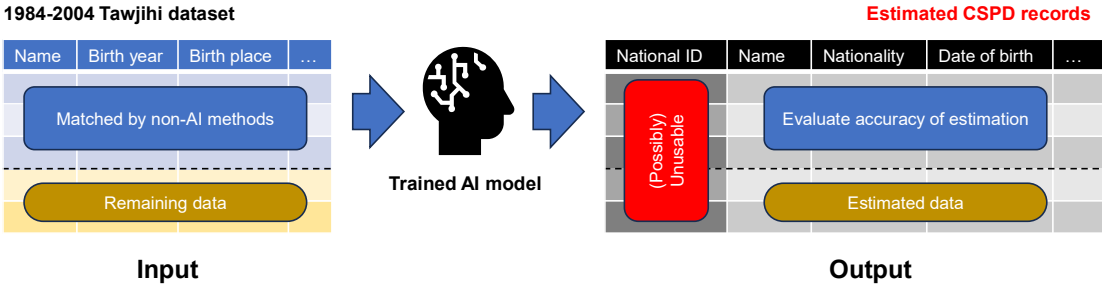

Figure 8: Evaluate performance of AI model

Note 1: It is expected that the estimated National ID number doesn't make sense at all because there is no logical or empirical way to estimate National ID numbers from such data as Name, Birth year, etc. of the students. It might be possible to get some tendencies of National ID numbers (in case the number is allocated based on some rules), but anyway the estimated National ID numbers by AI are just unusable because the National ID numbers must be exact and 100% accurate with no error (even 1 digit difference is unacceptable).

Note 2: However, it might be possible for the AI model to estimate other columns in CSPD data by learning how to correct inconsistent or misspelled writing of names, addresses, etc. It might also be possible to estimate better matching with other type of data such as higher education data. So the applicant should try at least one dataset other than CSPD to examine the matching performance and compare it to CSPD case.

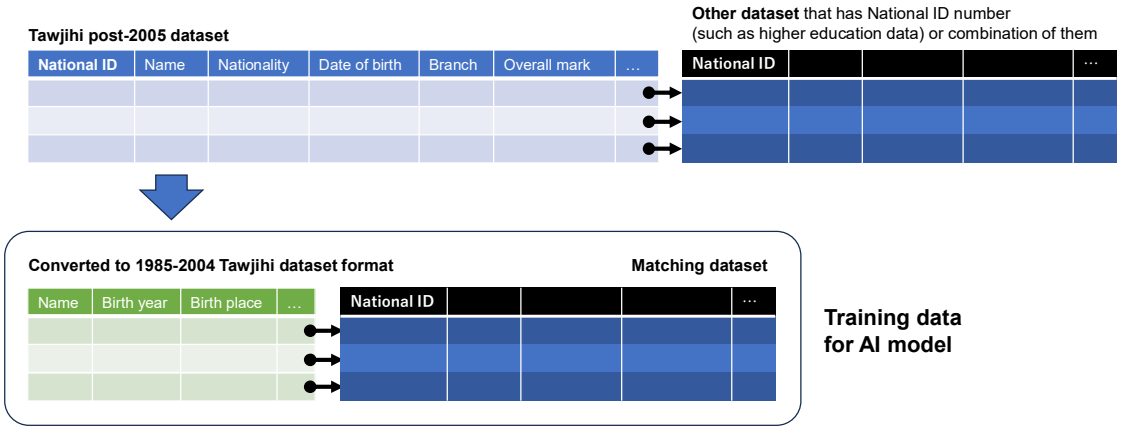

Figure 9: Use different dataset to measure the matching performance

(5) Key Performance Indicators (KPI)

- The percentage of recovering National ID number by using multiple non-AI methods must be significantly better than applying single non-AI method described in the example (82%)
- The accuracy of estimation of matching CSPD record (for remaining data) by AI methods should be 90% or above

(6) General conditions of Use case 1

- In this PoC, the applicant can exclude data with no assigned National ID number (such as non-Jordanian data who has only passport numbers that may change from time to time, or data with serial numbers that are not National ID numbers).

(7) Expected output (deliverable) of Use case 1

- A subset of 1985-2004 Tawjihi dataset that contains determined National ID number (by non-AI methods). This means that the subset should have additional data column for the determined National ID number.
- A working script or program (with source code) of non-AI methods that are applied to the dataset.
- A subset of 1985-2004 Tawjihi dataset that could not determine National ID number (by non-AI methods) but estimated it by AI method. This means that the subset should have additional data column for the estimated National ID number.
- A technical report describing both the non-AI methods and AI methods that are applied to

the dataset.

- The non-AI part of the report should contain the list of applied methods (algorithms), technical explanation of each method, sequence of applying these methods, and result of each application of methods (number of recovered data and its percentage (relative to the original data size) for each method).
- The AI part of the report should contain the technical explanation of chosen AI model, preparation of training data (including used datasets other than CSPD), result of applying AI models, and their evaluation.

Use case 2: Linking municipality property data to DLS data

(1) Background

Department of Lands and Survey (DLS) is the official authority responsible for data and property documentation, both for apartments and land. DLS stores its data in an Oracle DB that includes three main parts:

- Plot / apartment data
- Owner data
- Ownership data

DLS has been collecting these data since 1929 and started data automation in 1986 in cooperation with the Royal Scientific Society. The data volume is large, with over 4,000 tables, most of which affect the "white page," the main page for a property. Over 100 transactions affect this data, such as inheritance, partitions, state properties, exemptions, exchanges, and other transactions. Not all information and processes have been automated yet. The department began entering information into the database system in 2003. Initially, a sequential number was used to identify each land data, but now a unique national number is used for the identification. Currently, up to 97% of lands are linked to a national number. There are around 6.1 million records in DLS database and the number is increasing as new land and apartment are added.

On the other hand, each municipality (under the Ministry of Local Administration: MOLA) has been managing its own land data and they use different identification number system from DLS, and there has been no linkage between the DLS data and municipality data before. There are around 1.2 million data in MOLA database and the number is increasing as new land and apartment are added.

Since DLS is considered as the authoritative primary source for property information, all government institutions should align and coordinate their information with the data held by the DLS. Therefore, DLS has been working to connect with municipality data through GSB (Government Service Bus) where municipalities can obtain the main land key (DLS Key) from

DLS database to uniquely identify the property data. The DLS key is a 15-digit number that consists of following parts.

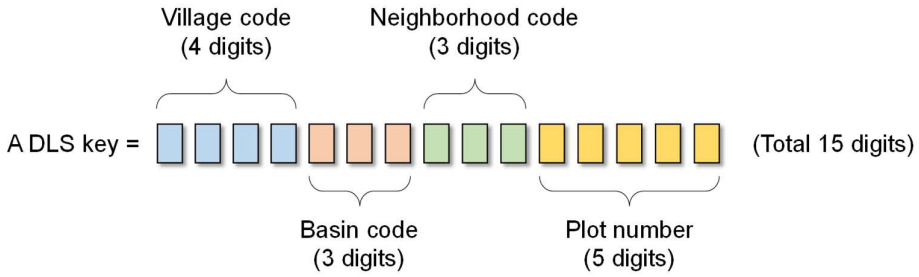

Figure 10: Structure of a DLS key

DLS encourages each municipality to store this DLS key into their property database so that data in MOLA database can link to data in DLS database, but the utilization rate of DLS Key is still low.

(2) Problems of linking municipality (MOLA) data to DLS

- Few municipality data have reference to DLS data (via DLS-key).
- Though both database have the same columns for village, basin, neighborhood, and plot numbers, they don't match well (see (4) below).
- MOLA database does not have strict data type checking that arbitrary text can be input into numeric fields (see (4) below).
- There are unregistered structures in remote areas that are not known to DLS.
- Names of government properties often have multiple names and are not unique in the database. There is an initiative underway to assign national numbers to government entities (Out of scope of this PoC).

(3) Goal of Use case 2

- **Goal 1**: Determine as many DLS key in MOLA database as possible by cross-referencing the MOLA data with DLS database using deterministic non-AI methods.

  Note: MoDEE data center has already experimented a matching method as follows.

  - Match village, basin, neighborhood, and plot numbers between MOLA and DLS data. This achieved the matching of nearly 960,000 records among 1.2 million records of MOLA dataset (matching ratio = 79.5%). There are still 20% of MOLA data that could not be matched with DLS data. Therefore, the Goal 1 is to increase the matching ratio as much as possible by using non-AI methods.
  - Since this result does not take data problems mentioned above into consideration, it could achieve better results if we carefully preprocess or cleans MOLA data, or if we try to do

other matching methods / algorithms.

- **Goal 2**: Apply AI method to estimate matching DLS records for MOLA dataset by training the AI model with MOLA records that already have determined DLS keys, and then apply the trained AI model to MOLA records that don't have DLS keys. Evaluate the AI model performance.

(4) Procedures to achieve goals

For Goal 1:

1. Examine the MOLA dataset and confirm / analyze the problems in the dataset. For example, we have already identified some problems of the MOLA dataset as follows:

   - The Plot number column in MOLA dataset ("`UNIT_NO_PERM`") contains non-numeric values such as the followings (it should be a number).

   Table 3: Non-numeric values found in UNIT_NO_PERM column of MOLA dataset

Non-numeric value	Example values	Possible solution
Blank		Replace with "0"
Number literals	00, 0000, 023, 046	Convert to numbers
Combined numbers	1+2	Divide the record into 2?
Sub numbers	27/1, 35/2, 230/101	Manual matching?
Unnecessary characters	*25, 1793., 633.	Remove unnecessary chars
Fraction numbers	1.11, 2.9, 5.9	Manual matching?
Arabic characters	مؤقث 37, ز, ص	Manual matching?
Others	12|||||||4, ., \	Manual matching?

2. Cleans the MOLA dataset based on the result of step 1. Note that no "data estimation" or "data augmentation" should be introduced at this step unless otherwise discussed and agreed with MoDEE and DLS.

3. Elaborate possible non-AI data matching methods (algorithms) to be applied to the MOLA dataset. All methods must be deterministic (no estimation) and must logically ensure 100% accuracy of the matched records. Typically, a non-AI method might be a computer algorithm (script) to be applied to the dataset by using data manipulations such as SQL.

   Important Note: If geographical coordinate data of plot is available in both dataset, it is possible to match records by using "spatial SQL"[2] calculation like shown in the figure below.

[2] https://en.wikipedia.org/wiki/Spatial_database

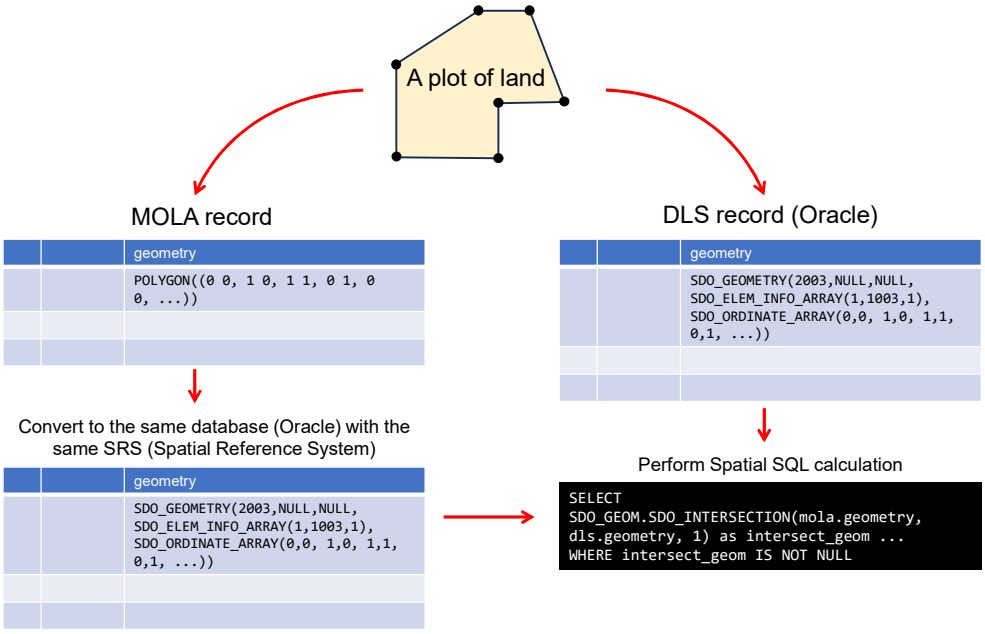

Figure 11: Using spatial SQL to match plots between MOLA data and DLS data

For example, for each record in MOLA dataset, you can calculate non-NULL intersection geometry with DLS database, which means the two records have plots at the overlapping (i.e. the same) location. Or you can first calculate CENTROID of all plots in both databases, then find the matching DLS record of a MOLA plot which has the minimum distance between centroids. All these calculations can be done by rather simple Spatial SQL commands. As far as both database have correct coordinate data, this Spatial SQL method should match almost all plots and there will be no need for using AI methods.

4. It is expected that single non-AI method might not recover all data (depending on the algorithms), so multiple methods should be applied subsequently to remaining data portion that could not be recovered by the previous method through "trial and error" manner as shown in the flowchart in Figure 3.

[Example of possible sequence of matching methods]

i. Cleans the plot number column of MOLA database.
ii. Match both database by using all 4 elements (village, basin, neighborhood, plot number) → Set aside the matched records.
iii. Match remaining data by using 3 elements (village, basin, neighborhood) and extract records that have only one matching record in DLS → Set aside the matched records.
iv. Match remaining data by using 2 elements (village, basin) and extract records that have only one matching record in DLS → Set aside the matched records.
v. Match remaining data by other elements (such as area of the plot, owner of the plot,

geographical location of the plot, shape of the plot, etc.). These operations may need a script or program using SQL to find matching record one by one like shown below.

> (Pseudo program)
> For each record in unmatched records of MOLA dataset:
> Find a record in DLS dataset where its village, basin, neighborhood are the same as MOLA record AND the difference of the area of plot is minimum.

For Goal 2:

1. Train an AI model using MOLA data that have already been matched with DLS.
   Note that village, basin, neighborhood, and plot number are obvious parameters that can directly link both datasets in the training dataset, but the remaining data for AI to estimate is those we cannot solve the linkage using these parameters. This means the AI model trained by these obvious parameters might not work well. In order to get better result, the applicant should consider training the AI model by not using all 4 obvious parameters (omitting Plot numbers, for example), try to use as many other parameters as possible, and try to use other datasets (owner data, for example) that may be relevant to the land data.

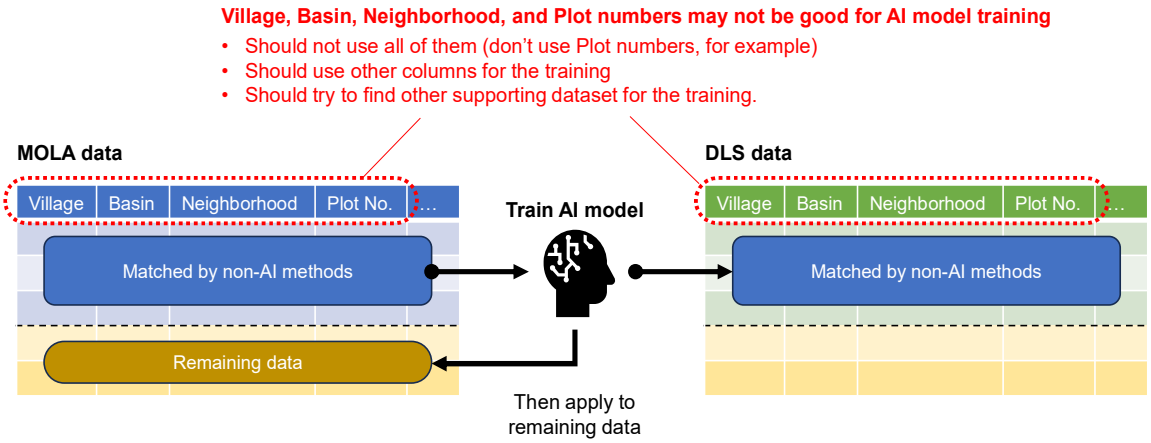

Figure 12: Train an AI model for MOLA-DLS data matching

2. Apply trained AI model to the remaining data that could not match with DLS data, and verify the performance by sampling some results manually and evaluate them. Note that we cannot apply AI model to the matched dataset because it's used for the training.

(5) Key Performance Indicators (KPI)

- The percentage of successful determination of DLS key in MOLA database using non-AI methods must be significantly better than the experiment done by MoDEE (80%).
- The accuracy of estimation of DLS key (for remaining data) by AI methods should be 90%

or above.

(6) General conditions of Use case 2

- In order to perform spatial SQL operation, you need to convert geometry data of both databases into the same RDBMS. Since Oracle used in DLS is not a free database, some open source RDBMS with spatial functions such as PostgreSQL + PostGIS[3] should be used in this use case.

(7) Expected output (deliverable) of Use case 2

- A subset of MOLA dataset that contains determined DLS key (by non-AI methods). This means that the subset should have additional data column for the determined DLS key.
- A working script or program (with source code) of non-AI methods that are applied to the dataset.
- A subset of MOLA dataset that could not determine DLS key by non-AI methods but applied AI method to estimate the DLS key. This means that the subset should have additional data column for the estimated DLS key.
- A technical report describing both the non-AI methods and AI methods that are applied to the dataset.
  - The non-AI part of the report should contain the list of applied methods (algorithms), technical explanation of each method, sequence of applying these methods, and result of each application of methods (number of recovered data and its percentage (relative to the original data size) for each method).
  - The AI part of the report should contain the technical explanation of chosen AI model, preparation of training data (including used datasets other than MOLA or DLS), result of applying AI models, and their evaluation.

Use case 3: Anonymizing Court decisions

(1) Background

Final court decisions are stored as document file format which contain summaries of statements and rulings. Currently, these court decisions are used by lawyers and legal services such as Qistas[4] and Qarark[5], and lawyers in charge of cases can obtain original (unmasked) court decision data directly from the court. With the implementation of the Personal Data Protection Law, personal information such as names, national IDs, addresses, or any other identifiers must be anonymized

[3] https://postgis.net/

[4] https://qistas.com/

[5] https://www.qarark.com/

from these files and replaced with appropriate substitutes, such as "Party A" instead of a full name, without compromising the context of case's decision. Currently, Qistas already provides manually anonymized files on its website. MoDEE data center plans to provide anonymization service of the court decision as a Web service in the future, and this PoC tries to implement its core AI-based anonymization software.

The number of court decision files are large.

- There are around 6 million historical court decision files.
- Around 40,000 files are newly added every month (around 0.5 million files per year)

Not all files are actually viewed by lawyers or potential service users, and the frequency of requests for the anonymized file might not be so often. Therefore, it is considered that the PoC prototype does not need to anonymize all files at once, but rather should dynamically generate anonymized file on demand according to the user's request. This mechanism has another advantage that we can always apply the latest anonymization AI model, and there is no need to replace all anonymized files again. It has also been agreed with the Ministry of Justice that the user of PoC prototype should be able to adjust the level of anonymization. The figure below shows the scope of this PoC and the overall structure of possible future anonymization service for court decisions at MoDEE.

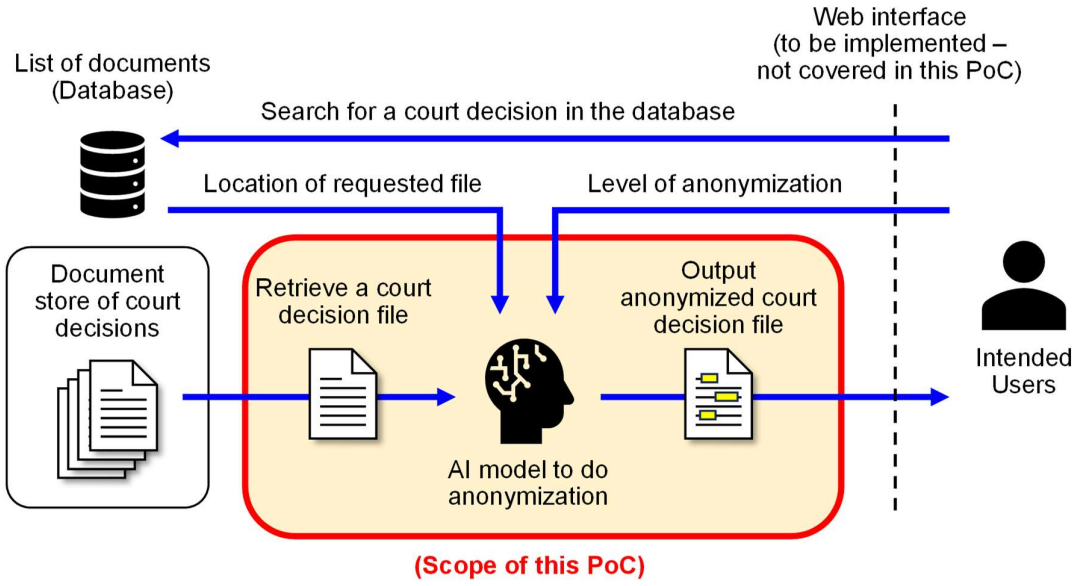

Figure 13: Structure of possible Web service for anonymization of court decisions in the future

Note that currently there is no actual plan of developing such a Web service at MoDEE, and this PoC would just try to prove the concept of court decision anonymization by using AI.

(2) Problem of the dataset (court decision files)

- Target data is not stored in a database but is stored as individual files.
- Human names can have up to 4 parts but sometimes the name is used in full form or in shorter form within the same document.
- Currently there are very few training data for AI (around 10 pairs of original files and their anonymized files) available and it needs to prepare more training data for AI.

(3) Goal of Use case 3

- Implement a console program (on Linux or Windows) which accepts three command-line parameters shown below, and then outputs an anonymized document file (in Microsoft Word format) of the input file.
  1. Path (including file name) to a court decision file to be anonymized (in Microsoft Word format). The program must not alter the input file, but just read the file.
  2. Path (including file name) to the anonymized (output) file. If omitted, output file will be created at the current path with the same file name as input file plus "_anon".
     (Example): **decision001.docx → decision001_anon.docx**
  3. Level of anonymization (in percentage: 0-100 as integer)
     where 0 means no anonymization at all, and 100 means full anonymization. The specified level will be used to set the threshold of anonymization as follows.

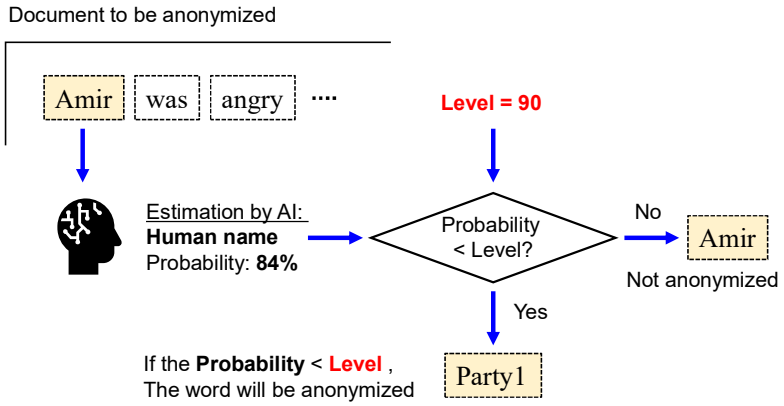

Figure 14: How level of anonymization works

Note: The AI model would always output the same result (with probability of each word in the document) regardless of the specified level value, and the level will be used for thresholding whether to anonymize the word or not based on the probability only.

- Anonymization must be applied to at least the following 4 kinds of elements in the court decisions. These are the mandatory elements that must be anonymized in this PoC.
  - ➢ Names of person or organization

- Addresses of person or organization
- Telephone number of person or organization
- National ID of person, company registration number, tax identification number, or any other numbers that can directly or indirectly identify a person or organization

Legally, there are more elements that should be anonymized for the purpose of full legal application as shown in the list below. These elements except for those listed above are not mandatory for anonymization in this PoC, but the applicant can consider anonymizing some of them in addition to the mandatory elements.

[Full list of anonymization target elements (optional in this PoC)]

1. Names of:
    a) The Accused
    b) The Claimant
    c) The Defendant
    d) The Accuser
    e) The Criminal
    f) The Appellant
    g) The Appellee
    h) The Petitioner (Court of Cassation)
    i) The Respondent (Court of Cassation)
    j) Suspect(s)
    k) Witness(es)
    l) Expert(s)
    m) Coroner
    n) Bailiff
2. National ID Number
3. Phone Number
4. Employment Number
5. Security Number
6. Home Address
7. Names of Cities/Towns
8. Place of Work
9. Role/Job
10. Land Specifications (Plot Number, Agricultural/Industrial Unit, etc.)
11. Name of Company, Organization, Shops, Unions
12. Company Registration Number(s)

There are also some elements that should not be anonymized. The applicant should exclude

following elements from the anonymization process.

[Elements that should not be anonymized]

1. "The Hashemite Kingdom of Jordan"
2. "His Highness King Abdullah II of Jordan"
3. Names of Judges
4. Names of Lawyers, acting in their capacity as lawyers.

- Target elements to be anonymized must be substituted by symbolic words (in Arabic) with number that distinguishes multiple identities (Pseudonymization). All symbolic words must be consistent throughout the document to preserve the accurate judicial / logical context. The Arabic words to be used for the substitution must be discussed and agreed with MoDEE and MOJ. Anonymized words must be clearly visible in the Word document by applying a predefined text format such as changing text color (or background color), etc.
  - Examples (in English for illustration purpose only)
    "Alice stole money from Bob" → "Party1 stole money from Party2"
    "Husni Soubar St 3, Amman" → "Address1"

  Note: The name of city / town can be kept not anonymized like Amman, Zarqa, etc.

(4) Procedures to achieve the goal

1. Choose several court decision files by sampling and analyze characteristics of these files. Perform statistical analysis of all court decision files to determine basic file characteristics such as the followings.

   - File format (how many files of older Word format (.doc), newer Word format (.docx), and other formats (if any))
   - File size (minimum, maximum, average)

2. Prepare for training data for AI model by manually anonymizing court decision files. There are already around 10 manually anonymized files at MoDEE, but the number of training data must be large enough for better training of AI model. It is expected that at least 100 or more training data should be prepared.

   Note 1: In order to prepare for many training data, it is recommended to use an annotation tool such as doccano[6].

   Note 2: In case the applicant plans to use publicly available Named-Entity Recognition (NER) model, the preparation of training data might not be required, but the applicant should

[6] https://github.com/doccano/doccano

assess the performance of ready-made NER, and retrain the model if needed.

3. Design and train an AI model to perform the anonymization. Applicant should propose the strategy to design the AI model. Some possible design strategies are:

   - Using Arabic Named-Entity Recognition (NER) tool or Morphological Analysis tool (if any) to detect all elements to be anonymized, then replace all detected elements with numbered symbols.
   - Using customized (cloned and locally-run) Arabic LLM by re-training it with training dataset to detect all elements to be anonymized, then replace all detected elements with numbered symbols.
   - It is better for the AI model to work on plain text data rather than Microsoft Word data.

   Note: It is not recommended to create a new AI model from scratch because it would take huge time and efforts including the preparation of large amount of training data.

4. Develop a console program that utilizes the trained AI model and perform necessary file handling, data format conversion (if any), etc.

(5) Key Performance Indicators (KPI)

- Accuracy of correctly detecting names of persons and organizations

Ground truth \ Detected by AI	Detected as Name (Positive)	Detected as non-Name (Negative)
It is a Name	True Positive (TP): > 95%	False Negative (FN): <5%
It is not a Name	False Positive (FP): < 5%	True Negative (TN): 95%

- Accuracy of correctly detecting addresses

Ground truth \ Detected by AI	Detected as Address (Positive)	Detected as non-Address (Negative)
It is an Address	True Positive (TP): > 95%	False Negative (FN): <5%
It is not an Address	False Positive (FP): < 5%	True Negative (TN): 95%

- Accuracy of correctly detecting telephone numbers and national ID numbers

Ground truth \ Detected by AI	Detected as Number (Positive)	Detected as non-Number (Negative)
It is a Number	True Positive (TP): > 98%	False Negative (FN): <2%
It is not a Number	False Positive (FP): < 2%	True Negative (TN): 98%

(6) General conditions of use case 3

- Investigative and juvenile cases are excluded in this use case as the Ministry of Justice

considers these data to be confidential.

- Use of open source software is allowed if the software is installed and run in the private environment (on premise) so that no data will be exposed to outside of MoDEE data center.

(7) Expected output (deliverable) of Use case 3

- A console program (and its source code) with a trained AI model that accepts an input file of court decision and outputs anonymized file as described in the Goal of Use case 3.
- All training data (pairs of court decision files and their anonymized files)
- A technical report describing the methodology used in the program, selection criteria of AI model, detailed result and explanation of anonymization performance.

3. Implementation body of the PoC program (Important)

This PoC program shall be implemented by a Public-Private-Academia consortium consisting of members from the target government organization (beneficiary of the PoC, and provider of data to be used in the AI system), private IT company (developer of the PoC solution), and academic AI researcher (the technical advisor in the field of the latest AI technology). The reason for formulating the three-parties consortium is to demonstrate the importance of collaboration among government, industry and academia for accelerating the development of local AI industry which contributes to solving socio-economic problems in Jordan. the focus of JICA project is to try to develop and enhance the capacity of local AI industry and academia instead of relying on foreign companies so that it would also contribute to generating local employment in the field of advanced technologies like AI. There is another reason to employ three-parties collaboration in this PoC that such collaboration has been proved to be very successful in Japan.

In this PoC program, the target government organization is MoDEE data center as primary stakeholder, plus MOE (Ministry of Education), DLS (Department of Land and Survey), MOJ (Ministry of Justice) as original data owners for each use cases. A representative person will be appointed in each government organization for this PoC and those persons will be members of the consortium. Private IT company and Academic AI researcher will be selected by a tender process as a joint team based on this ToR. This means that the applying party must not be an IT company alone or researcher alone, but must be a joint team of both. All tender processes will be implemented by the Project Secretariat together with JICA consultant team.

The expected roles of the three parties are summarized in the table below.

Table 4: Expected roles of PoC consortium members

Roles	Government Organization	Private IT company	Academic AI researcher
Basic role	Beneficiary of solution, Provider of data	Developer of AI solution	Research and Advisor on the AI method to apply in PoC

Roles	Government Organization	Private IT company	Academic AI researcher
Reporting	• Communicate with both private IT company and academic AI researcher	• Communicate with both government organization and academic AI researcher.	• Communicate with both government organization and private IT company
		• Must have a unified contact point for reporting to and communicating with stakeholders (see chapter 4 below)	
Designing	• Provide input as the initiator of problem to be solved. • Assess availability of data to be used in the solution. • Provide sample data for designing the solution.	• Collaborate as a team to do the followings. • Interview government organization for requirements • Design prototype PoC solution based on the result of interview as well as information on available data • Validate the system design and its required data	
Implementation	• Provide full data that is necessary to build / train AI model.	• Collaborate as a team to do the followings. • Develop a prototype working solution for the PoC	
Testing	• Evaluate the result of testing and provide advice from the standpoint of data owner	• Perform testing of the PoC solution and solve issues found in the testing	• Verify and validate the test result and provide technical advice on the improvement of AI model from the standpoint of advanced AI researcher
Evaluation	• Evaluate the performance by comparing the result with human	• Evaluate the result from the standpoint of implementation methods	• Evaluate the result from the standpoint of performance of applied AI technologies

4. Unified Contact Point of the Joint Team (Focal point)

The Joint Team must have a single, unified point of contact (focal point) to reporting to and communicate with stakeholders (MoDEE, AI Steering Committee, JICA consultant team, target government organization) which represents both private IT company and academic AI researcher. Any coordination among the consortium members must be done internally within the consortium, and each member must not communicate independently or directly with the stakeholders. Stakeholders will not provide any coordination within the consortium, but it would be possible for the unified point of contact (focal point) to consult such matters with stakeholders.

5. Requirements for private IT company

a. Must be a Jordanian company registered with the CCD with relevant licensing segmentation to provide technology services. A valid licensing permit must be present.
b. If the applicant is a consortium of Jordanian companies, all must be registered with the CCD and have a valid licensing permit. Should the consortium include foreign companies, they should not be consortium leaders nor can they render more than 20% of work required.

c. Must have a major IT development team in Jordan that consists of 80% or more team members who are Jordanian nationals or permanent residents in Jordan.
d. Experience in development of systems that employ machine learning technologies and other AI-related technologies.
e. Minimum two (2) years in system development business.
f. Experience in the following fields is a plus.
   - Complex RDBMS operation using advanced SQL
   - Spatial RDBMS and Spatial SQL
g. Must be able to work with stakeholders from public sectors and academia.
h. Must be able to form and work as a qualified team of IT engineers and academic AI researcher(s).
i. Must perform roles described in Table 1 at the column of "Private IT company".
j. Must have audited financial statement for the last two (2) years.
k. Must commit to the ethics of artificial intelligence (Jordan AI code of Ethics)

6. Requirements for academic AI researcher

a. Must be a researcher or a professor at a university or a research institute in Jordan in the field of AI.
b. Minimum five (5) years' experience as a researcher with three (3) years in the field of academic research of AI technology.
c. Experiences in Arabic NLP (Natural Language Processing) research is required
d. Must be able to work with stakeholders from public sectors and IT industry.
e. Must be able to form and work as a qualified team with IT engineers of private company.
f. Must perform roles described in Table 1 at the column of "Academic AI researcher".
g. Must commit to the ethics of artificial intelligence (Jordan AI code of Ethics)

7. Tasks of the joint team

The Joint Team should perform following tasks:

a. Review on MoDEE's AI Strategy and Implementation Plan (2023-2027) and the Work Plan of JICA project (provided separately) to gain understanding of the background of the PoC program.
b. Hold a kick-off meeting of the PoC consortium consisting of the Joint Team and representatives from target government organization (beneficiary of the PoC program) to discuss and confirm the content and schedule of the development of PoC solution.
c. Participate in meetings related to the implementation of the PoC program with sub-committee of the AI Steering Committee as well as JICA project team members.
d. Perform tasks of three data preprocessing / cleansing use cases described in 2.
e. Report the progress of PoC program to AI Steering Committee on regular basis (bi-weekly).
f. Submit all deliverables specified in 2.

g. At the end of the work, write a fully comprehensive completion report and submit the report to the AI Steering Committee.

8. Deliverables

The Joint Team should submit the following deliverables:

a. All deliverables described in 3 use cases in chapter 2.
b. Bi-weekly progress reports. During the test-run / evaluation period, the reports should contain the current performance (KPI) of each use case.
c. Fully comprehensive completion report that includes the followings.
   - Summary of development activities with schedule
   - Summary of evaluation result of each use case (KPIs)
   - Lessons learned and recommendations for areas to improve for possible full-scoped project in future.

9. Duration and Timeline

The duration of this PoC program is six (6) months from the beginning of assignment. The expected timeline of PoC program is shown in the table below.

Table 5: Expected timeline of the PoC program

Activity	Month					
	1	2	3	4	5	6
Use case 1: Determining national ID number in Tawjihi Data (1985-2004)						
1. Examine 1985-2004 Tawjihi dataset						
2. Cleans the dataset based on the result of 1.						
3. Elaborate possible non-AI methods						
4. Implement non-AI methods						
5. Design and train AI model						
6. Apply and evaluate AI model						
7. Make and submit deliverables						
Use case 2: Linking municipality land data to DLS data						
1. Examine MOLA dataset						
2. Cleans the dataset based on the result of 1.						
3. Elaborate possible non-AI methods						
4. Implement non-AI methods						
5. Design and train AI model						
6. Apply and evaluate AI model						
7. Make and submit deliverables						
Use case 3: Anonymization of court decisions						
1. Analyze court decision files						
2. Research and determine AI model						
3. Prepare for training dataset						
4. Train (Fine-tune existing) AI model						
5. Develop a console program						
6. Make and submit deliverables						
Create final report and submit all deliverables						

Note: When the applicant makes quotation, the cost of remuneration for required human resources must be appropriately calculated by clearly and accurately estimate their efforts.

10. Confidentiality and Intellectual Property

a. Joint Team should respect the confidentiality of shared information and should agree on the handling of intellectual property rights as outlined in a separate agreement. Some data in this PoC such as court decisions are very sensitive data that the applicant might have to sign NDA with the relevant stakeholders.

b. Intellectual Property of the developed PoC solution should belong to the Joint Team. The beneficiary (MoDEE, MOE, DLS, MOJ) should have the right to use the PoC solution for unlimited time.

11. Budget and Important Contract Conditions

a. Budget for the PoC program will be allocated, managed, and disbursed by JICA according to the JICA's procurement rule.
b. The contract of this PoC program will be a sub-contract of JICA through JICA consultant team of Japan Development Service Co., Ltd.
c. There will be no Jordanian government involving in the contract.
d. The law that governs the contract will be the law of Japan.
e. There will be no advance payment.

12. Gender consideration

JICA has a global agenda and strategy for Gender Equality and Women's Empowerment. In this context, JICA plans to give additional appreciation to the female participants in the PoC when evaluating the proposal in tender process.