Legacy Data Cleansing Q&A

Date: 15-4-2025

Question No.	Question	Answer
Q1	Could you provide a complete inventory of all data sources (databases, applications, files) that need to be cleansed and migrated?	The data is hosted in a centralized structured database at MOHE, collected from all public and private universities.
Q2	Are there documented data dictionaries, entity-relationship diagrams, or database schemas available for the legacy systems? If yes, when can we access them?	These documents in details will be provided to the winning bidder.
Q3	Is there a target data model or schema already defined for the post-cleansing environment, or will we need to design one as part of this project?	The winning bidder will be responsible for designing the post-cleansing data model.
Q4	Could you please confirm the nature of the legacy data? Specifically, is it primarily structured (e.g., databases, spreadsheets), semi-structured (e.g., XML, JSON), or unstructured (e.g., scanned records, images, handwritten files)?	The legacy data is structured.
Q5	What is the total estimated volume of data to be processed (in GB/TB or record count) across all institutions? Could you provide a breakdown by university and data type?	Database size is 7.5 GB witch include all the institutions
Q6	When referring to "educational institutions," does the scope include public universities only, or are private institutions also part of the project?	The scope includes both public and private universities.
Q7	What is the time span of the legacy data (e.g., records from 2000 to present)? Are there specific retention requirements for different data categories?	Dump size 7.5 GB approximately with more than 32 million records
Q8	Approximately how many distinct tables/entities and attributes/fields are involved across all systems? Are there complex relationships between entities that need to be preserved?	Details will be provided to the winning bidder.

Q9	Have any preliminary data quality assessments been conducted? If yes, could you share the findings regarding the most common or critical data quality issues?	No preliminary assessments have been conducted.
Q10	Is there consistency in data definitions and formats across different universities, or does each institution have its own data standards and definitions?	Yes, all universities follow a standardized format and structure.
Q11	What percentage of the data contains personally identifiable information (PII) or other sensitive data requiring special handling?	All data is personal and requires special handling.
Q12	Are there known data quality issues specific to particular universities or data types that will require specialized cleansing approaches?	All data is in a similar format with the same quality issues, though the severity may vary.
Q13	Is there a preference for how data cleansing should be implemented—through ETL processes, in-database transformations, specialized data quality tools, or a combination of approaches?	Any preferred approach is acceptable.
Q14	Should all profiling and cleansing steps (e.g., null checks, format standardization, deduplication, etc.) be applied uniformly across all datasets, or will the process vary depending on the data source/type?	All data is stored in a single database hosted at MOHE. You are free to use a uniform or varied approach depending on your methodology.
Q15	For the data quality dashboard requirement, do you have specific key performance indicators (KPIs) or metrics that must be tracked and visualized?	Yes and more details will be provided to the winning bidder: 1. Data Accuracy 2. Data Completeness 3. Data Consistency 4. Data Timeliness 5. Data Integrity 6. Data Usage and Monitoring
Q16	What are the expected success criteria for data cleansing? Is there a minimum acceptable threshold for data quality improvement?	95% data quality for Jordanian records and 70% for non-Jordanian records after 2011.

Q17	Will we have direct access to the source systems, or will data extracts be provided? Will we need to work within your network environment, or can processing be done offsite?	A copy of the data will be provided. All work must be done on-site; remote work is not permitted.
Q18	Are there specific technology constraints or preferences for the tools and platforms used in the data cleansing process (e.g., specific database platforms, ETL tools, or programming languages)?	There are no specific constraints or preferences.
Q19	For the student record integration mentioned in the RFP, could you elaborate on the CSPD & PSD integration requirements? What external systems need to be integrated, and what is the nature of these integrations?	The integration is for data cleansing purposes only.
Q20	Would you prefer a phased approach to data cleansing (e.g., university by university or data type by data type), or should all data be processed simultaneously?	You may choose the approach, but all data will be provided to MODEE simultaneously.
Q21	Would you prefer to receive all final deliverables as a single delivery at the end of the project, or would a phased delivery approach (e.g., per institution or per data phase) be more appropriate?	According to the Arabic sample agreement (page 16), 100% of the payment is issued after full project delivery.
Q22	Are there any specific universities or data types that should be prioritized in the cleansing process?	No prioritization is required.
Q23	Will the scope include creating data pipelines for ongoing data quality monitoring after the initial cleansing, or is this limited to a one-time cleansing effort?	Yes, the scope includes the creation of ongoing monitoring pipelines.
Q24	Are there existing data governance frameworks or policies we should align with? Who will be responsible for making decisions about data standardization and quality rules? Will we need to develop data governance documentation as part of our deliverables?	The winning bidder will be responsible for this.

Q25	Are there industry-specific or national data standards that must be followed for educational data in Jordan? If yes, could you provide references? Additionally, will the selected vendor be responsible for developing any new standards or metrics from scratch, in coordination with MoDEE?	The winning bidder will be responsible for compliance and/or development of necessary standards.
Q26	Who are the key stakeholders from each university that will be involved in the project? Will there be dedicated subject matter experts (SMEs) available to assist with data validation and business rule definition?	Refer to Q27. Coordination will be solely with MOHE and MODEE.
Q27	What level of coordination will be required between our team and the university IT departments? Will we have direct access to their technical staff?	There is no need for coordination with universities. Communication will only be with MOHE and MODEE.
Q28	Given the 180-day project timeline, are there any critical milestones or deadlines we should be aware of within this period?	All deliverables must be submitted within the 180-day timeframe.
Q29	Will MODEE provide any resources (personnel, infrastructure, tools) to support the project, or is the vendor expected to supply all necessary resources?	MODEE will provide workstations, virtual machines, database infrastructure, and licenses. The vendor is responsible for supplying the necessary tools.
Q30	Are there constraints on where the work must be performed? Can some tasks be performed remotely, or must all work be done on-site?	All tasks must be performed on-site. Remote work is not allowed.
Q31	Is there a defined budget range for this project that we should be aware of when preparing our proposal?	The budget range cannot be shared.
Q32	How should we structure our pricing proposal? Are you expecting a fixed price for the entire project, or would you consider a phased pricing approach based on data volume or complexity?	Refer to the Arabic sample agreement (page 16). The project is fixed-price, paid upon full delivery.
Q33	Are there known issues with the legacy systems that might complicate data extraction or processing (e.g., performance limitations, stability issues, or access restrictions)?	Known issues include data quality and integrity.

Q34	Have there been previous attempts to cleanse or migrate this data? If yes, what challenges were encountered, and what lessons were learned?	No previous attempts have been made.
Q35	What continuity plans are in place if the source systems need to remain operational during the data cleansing process?	A copy of the data will be provided. Updates will be reflected in the production system after solution acceptance.
Q36	Could you provide a detailed view of your current data schemas, table structures, and relationships (including primary/foreign key relationships) across the datasets?	a. There are no complex relationships between tables and we have the following constraints: b. TABLEs=85 c. COLUMNS = 1117 d. Constraints types: i. Primary key = 67 ii. Referential Integrity = 233 iii. Check = 1 iv. Unique key = 3
Q37	Are there any specific complexities in the relationships between tables that we should be aware of (e.g., many-to-many relationships)?	Related to Q36
Q38	What types of format anomalies have you encountered across datasets (e.g., inconsistent date formats or address formatting)?	Related to Q9 and Q12. Format inconsistencies exist and will be addressed during cleansing.
Q39	Are there any specify the most critical data fields that should always be complete and accurate?	Related to Q16. Priority fields should meet defined data quality thresholds.
Q40	Are there any specific criteria or rules that define "incomplete" or "inaccurate" data in your context?	Related to Q16. Cleansing criteria are defined by success thresholds: 95% (Jordanian), 70% (non-Jordanian).
Q41	Are there particular data points or outliers that you are concerned about? (e.g., GPA > 4.0)	Related to Q9 and Q12. Outliers will be addressed as part of anomaly detection.
Q42	Are there specific algorithms or thresholds that you would like us to implement for anomaly detection?	Related to Q16. Detection approaches can be proposed by the bidder.
Q43	Are there existing data constraints in place (e.g., primary/foreign key constraints) that we should enforce, or do you require us to implement these constraints during profiling?	Related to Q36

Q44	Should we perform data integrity checks to ensure that foreign key references and relationships are maintained consistently?	Related to Q8. Yes, integrity checks are expected.
Q45	Are there specific field relationships that we should focus on? (e.g., "Graduation Date" depends on "GPA")	Related to Q8 and Q16. Yes, logical dependencies will need validation.
Q46	What attributes should be used to identify duplicate records? Do you have a preferred method (e.g., merging or flagging)?	Related to Q13. Deduplication approach can be determined by bidder.
Q47	Are there any specific field types or structures that require standardization (e.g., postal codes)?	Related to Q14. Yes, standardization is expected across fields.
Q48	Do you have any predefined rules or external data sources to validate values?	Related to Q24. Yes, external validation may be incorporated.
Q49	Are there any fields that should undergo correction based on business rules or external datasets?	Related to Q24. Yes, this may include fields like student ID or address.
Q50	Do you want the deduplication process to be fully automated or manually reviewed?	Related to Q46. Can be automated or involve manual review.
Q51	Are there specific logical checks or validation rules we should apply?	Related to Q16 and Q45. Yes, logic checks are expected.
Q52	Should we apply consistency checks (e.g., graduation date matches credits earned)?	Related to Q45 and Q16. Yes, consistency validation is required.
Q53	Are there any legacy data fields that need to be mapped to modern standards?	Related to Q3. Yes, mapping is expected.
Q54	Should we ensure the compatibility of old data with your current system's data structure?	Related to Q3. Yes, compatibility is required.
Q55	Do you require us to apply normalization techniques (e.g., 3NF)?	Related to Q3 and Q8. Yes, normalization may be needed.
Q56	Are there specific entities or relationships where normalization is most critical?	Related to Q8. To be determined with the winning bidder.
Q57	Do you plan to integrate external data sources to supplement legacy records?	Related to Q19. Yes, integration with CSPD & PSD.
Q58	Should we augment missing data (e.g., emergency contacts) with external databases?	Related to Q57. Yes, where appropriate.
Q59	How should we handle incomplete data during transformations (e.g., missing GPA)?	Related to Q16. Strategy to be proposed by the bidder.

Q60	Do you have a set of data standards or metrics, or should we propose a framework?	Related to Q24 and Q25. The bidder should propose and apply standards.
Q61	Are there any rules for maintaining data consistency and accuracy?	Related to Q24. Yes, rules must be defined and applied.
Q62	Do you require documentation of the transformation/migration logic?	Related to Q24. Yes, full documentation is required.
Q63	Should we document the legacy structure and migration process for audits?	Related to Q24. Yes, for future reference.
Q64	How often should data quality audits be conducted post-migration?	Related to Q23. Frequency can be defined by the project.
Q65	What issues should be included in post-migration audits?	Related to Q23. Missing data, broken links, inconsistencies.
Q66	Are you interested in automated monitoring tools for data quality?	Related to Q23. Yes, automated quality checks are encouraged.
Q67	What specific key metrics should be in data quality reports?	Related to Q15. KPIs to be defined with Areej.
Q68	Should we track university-specific issues like missing data %?	Related to Q15. Yes, tracking across institutions is expected.
Q69	What KPIs should be on the dashboard for monitoring?	Related to Q15.
Q70	Should the dashboard allow drill-down by university or dataset?	Related to Q15. Yes,
Q71	Should we provide a technical document of algorithms used?	Related to Q24. Yes, detailed technical documentation is required.
Q72	Do you have a target percentage for data "cleanness"?	Related to Q16. 95% (Jordanian), 70% (non-Jordanian) post-2011.
Q73	Are there specific formats or APIs needed for data integration?	Related to Q19. Details will be provided if applicable.
Q74	What level of automation is expected during migration?	Related to Q19. To be defined with the winning bidder.
Q75	Should we set up automated systems for ongoing validation (e.g., Alteryx)?	Related to Q66 and Q23. Yes, automation is encouraged.
Q76	How many databases are expected to be analyzed?	1 Database
Q77	What is the estimated total number of database tables?	85 Table
Q78	To detect data entry errors, we require business rules and data standards. Can you provide these?	Will be provided to the winning bidder
Q79	Are there any specific tools that must be used for data profiling and cleansing?	No specific tools are mandated.
Q80	Are there any restrictions on using tools like Microsoft Studio for running ETL processes for data cleansing?	There are no preferences for specific tools. However, any proposed tools must be approved by the security team.

Q81	After project completion, should these tools be integrated into the environment for ongoing data cleansing?	Yes.
Q82	Should data cleansing be performed directly on the existing database?	No, cleansing should be performed on a copy of the database.
Q83	Or should a new database be created to store and manage the cleansed data moving forward?	This can be proposed by the bidder based on the overall solution design.
Q84	If feasible, can the project include addressing the root causes of data entry errors to minimize future issues, such as implementing validation rules at the data entry stage?	Yes, this is expected as part of the project.
Q85	Will the ministry provide a sample dataset or metadata description to better understand the current data structure?	No, not allowed
Q86	Are there any third-party systems the solution must integrate with (e.g., other MoDEE systems or university platforms)?	No, this is not part of the winning bidder's scope.
Q87	Will a sandbox or testing environment be made available to the winning bidder?	Yes.
Q88	Can open-source technologies be used as part of the data cleansing or dashboard solution?	Yes, as long as they are approved by the security team.
Q89	Could you please specify the total volume of legacy data (in terms of records, size, and number of institutions) that will need to be cleaned and migrated?	Related to Q7
Q90	Are there any predefined quality standards or metrics for the data that need to be followed during the cleansing process, or should the bidder propose them?	The bidder should propose the standards.
Q91	What types of legacy data sources (e.g., databases, spreadsheets, text files, etc.) are involved in the project? Will these sources be provided in a specific format?	All legacy data sources are structured databases.
Q92	Will there be a need for significant data transformation, such as converting old formats into new ones, or is the	Recommendations are welcome if transformation is necessary.

	primary focus on cleaning and standardizing data?	
Q93	Are there specific tools or technologies preferred or mandated by MoDEE for data profiling and cleansing, or is the bidder free to choose their tools?	The bidder is free to propose tools, but they must be approved by the security team.
Q94	Is the bidder expected to integrate the cleansed data into existing systems, and if so, which systems should be integrated with (e.g., MoDEE's existing databases or platforms)?	No, integration into existing systems is not required.
Q95	To what extent should the data cleansing process be automated? Is there a preference for using manual versus automated methods?	The process should be automated, at minimum using scheduled jobs.
Q96	Will there be a need for validation of the data cleansing algorithms? If so, how will the validation process be conducted, and who will be responsible for it?	Validation will be handled by the quality team at MOHE/MODEE.
Q97	Could you clarify the exact format and content of the "Data Quality Reports" for each university? Are specific reporting standards or templates expected?	All data is stored in a single database at MOHE. Reports should be generated at the database level, not per university.
Q98	Can you provide more details about the expected features and functionalities of the data quality dashboard? Will it be integrated with any other existing systems?	Related to Q15
Q99	Are there specific performance metrics or KPIs that bidders should be aware of in relation to project delivery and quality?	Related to Q48
Q100	If there are significant changes to the scope of work during the project, what is the process for handling these changes, and will additional costs be covered?	Yes, changes will be covered if necessary.
Q101	Could you please specify the number of universities that will be involved in this project, and if available, provide any details about the data from these institutions?	All universities are included. Their data resides in a single unified database hosted at MOHE.