A Proposal for the Data Quality Objectives Process to Apply to the Use of Artificial Intelligence for Getting Data on the Spread of Radioactive Contamination

Younghoon Oh[1], Younghee Park[1], Jiyon Lee[2] and Namhun Kim[1*]
[1]*Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology, Ulsan, 44919*
[2]*Korea Institute of Nuclear Safety, Republic of Korea, Daejeon, 34142*

**Corresponding author: nhkim@unist.ac.kr*

1. Introduction

In South Korea, there are 25 nuclear power plant generators operating as of March 2023 [1]. Although the radioactivity from the plants is kept safe, nuclear accidents like Chernobyl and Fukushima are always a possibility. Pollution dispersion surveys are used to gather data in such an emergency so that decisions about resident protection measures can be made. Compared to data acquired during regular times, emergency data has a greater range of quality. To determine whether the quality of the data is appropriate for the intended application, we must employ some sort of criterion. Our research is being undertaken to utilize Data Quality Objectives (DQO), a way to give a data collecting plan to enough amount and quality of data to support the objectives, designed by the U.S. Environmental Protection Agency (EPA), to solve these issues [2]. Current studies only offer solutions to the DQO process's current limits for analyzing the distribution of radioactive contamination; they do not take into account artificial intelligence (AI) technology's scalability, which is a growing trend worldwide [3]. This study will make use of the DQO process' benefits while also suggesting a scalable DQO method for use in AI applications.

2. Methodology

A schematic diagram of the DQO process is illustrated in Fig. 1, and the considerations for applying AI for each step are summarized in Table I. The main purpose of the DQO process and additional considerations for AI application are described step by step as follows.

The first step in the DQO process defines the problem so that the focus of the project is unambiguous, selecting the decision maker for the project, setting available resources and research deadlines. In the case of AI applications, this should include dataization of the problems and identification of data sources since it is necessary to clearly specify the problem to be solved by AI and transform it into a suitable form.

The second step is to specify a clear goal by defining the research questions and countermeasures based on the results of concern by answering the questions. In the case of AI, since conclusions are drawn based on the collected data, parameters, and the purpose of the model, it is necessary to set the direction of the model by clarifying the causal relationship of the problem. According to the goal of the study, model direction will be decided such as predictive model and classification model. This direction becomes another consideration of the input which is considered at the next step.

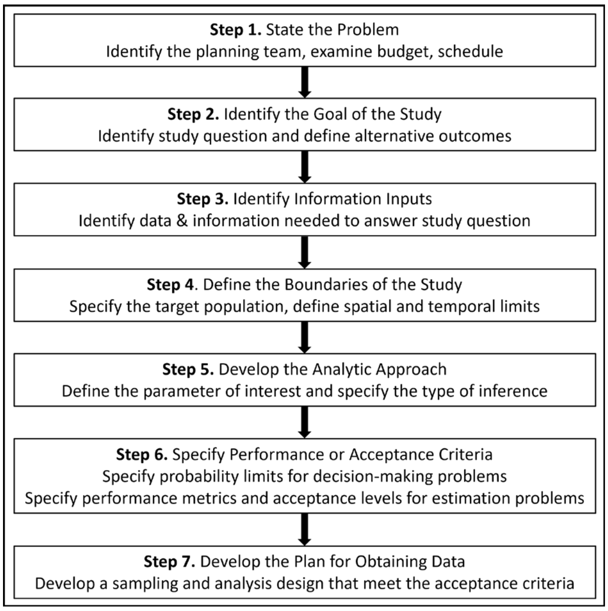
Step 1. State the Problem
Identify the planning team, examine budget, schedule
Step 2. Identify the Goal of the Study
Identify study question and define alternative outcomes
Step 3. Identify Information Inputs
Identify data & information needed to answer study question
Step 4. Define the Boundaries of the Study
Specify the target population, define spatial and temporal limits
Step 5. Develop the Analytic Approach
Define the parameter of interest and specify the type of inference
Step 6. Specify Performance or Acceptance Criteria
Specify probability limits for decision-making problems
Specify performance metrics and acceptance levels for estimation problems
Step 7. Develop the Plan for Obtaining Data
Develop a sampling and analysis design that meet the acceptance criteria

Fig. 1. The Data Quality Objective Process [2]

The third step is identifying information inputs. This is the step of analyzing and selecting necessary data to solve the decision-making and estimation problems defined in the previous steps and confirms the type and source of the essential information. In this process, it is required to determine the information: the types of variables that need to be collected, the types of information that meet performance or acceptance criteria, appropriate data for the direction of the AI model, and appropriate sampling and analysis methods. Based on the input data which is defined in this procedure, type of the AI model will be decided. Images data would lead the model to CNN based model and time series data lead it to RNN based model. After the decision, pre-processing and data correlation verification are necessary because the types, acceptance criteria, and types of variables directly affect the parameters of the AI models.

The fourth step is defining boundaries. This is the step to identify the population for research and to specify spatial and temporal features associated with the decision-making or estimation problem. Since it is necessary to prove that the data extracted from a specific area is representative of the population, using data correlation indicators to evaluate the quality of data which is classified by the defined boundaries. At this point, methods such as interquartile range need to be used to minimize inappropriate data such as outliers, thereby increasing representative of the population and quality of the data.

The fifth step involves the development of an analysis methodology, which includes identifying population parameters and determining an action level. These activities enable the planning team to draw meaningful conclusions from the data. Applying AI can involve exploring different approaches to data analysis and selecting the most appropriate one based on the data and goals of the AI application. For example, if researcher want to classify the degree of contamination through image classification, features (e.g., variance) between acquired datasets should also be considered in order to select a specific model such as UNet or ResNet among various CNN models.

The sixth step of the DQO process specifies performance or acceptance criteria, taking into account the fact that the data is not perfect and may contain errors. The criteria will help minimize the risk of misleading conclusions or exceeding acceptable levels of uncertainty. When applying AI, it is important to consider potential biases in data and algorithms. Because bias can occur from many sources, including sampling strategies, data collection processes, or algorithms used to analyze the data, performance or acceptance criteria should be set while considering criteria selection to minimize bias. To this end, if K-fold cross validation and regularization techniques are used, bias can be minimized even if the collected data is small.

The seventh step involves developing a resource-efficient design to collect data that meets the requirements specified in the previous steps. This design should be sufficient to achieve research objectives or to maximize the amount of information available within a fixed budget while achieving performance or acceptance criteria. In addition to the main activities of the step, it is possible to consider using AI models by utilizing data from existing similar situations for AI applications. In South Korea, AI Hub[4], Public data portal[5], etc. are disclosing AI models using data in various situations, so they can be reviewed together. AI models can be used to optimize sampling designs by analyzing large datasets and identifying patterns or correlations that can inform sampling strategies.

Table I: Considerations for AI applications

Step	Definition	Considerations for AI application
1	State the problem	Include dataization of the problem and identify the data source
2	Identify the goal of the study	Concern the direction of the model for causal relationships and goals
3	Identify information inputs	Require data correlation verification and data preprocessing
4	Defines the boundaries of the study	Carry out additional data processing while including indicators for evaluating the quality of data with boundaries
5	Develop the analytic approach	Choose the best approach based on the data and objectives of the AI application
6	Specify performance and acceptance criteria	Set criteria to account for potential bias in data and algorithms
7	Develop the plan for obtaining data	Optimize sampling and analysis designs by utilizing data analysis results from existing similar situations

3. Discussions and Conclusions

The DQO process differs depending on the problem type of utilizing research data, which is divided into decision-making and estimation. This study focused on suggesting additional considerations at each step for the applications of AI rather than mentioning the detailed problem types accordingly. In addition, since case verification has not been conducted, additional studies are needed. In the future, it will be necessary to confirm and verify the DQO step by step to secure high-quality data related to radioactive contamination. However, this study is significant as it introduces the DQO process and proposes a new methodology that incorporates AI analysis for evaluating the distribution of radioactive contamination, thereby addressing critical needs in the field. If a methodology is developed by applying the suggestions introduced in this study, the data extracted through this can be immediately analyzed by AI, ensuring efficiency and scalability. For the efficient establishment and operation of the information system required for radioactive impact assessment, collection, analysis, and management of measurement standards are required. The proposed evaluation methodology can be used as a guideline for the development of radioactivity measurement technology, and it can support the improvement of the data quality that is the basis for safety evaluation.

REFERENCES

[1] Korea Atomic Industrial Forum, https://www.kaif.or.kr/ko/?c=188&s=188

[2] U. States, “Guidance on Systematic Planning Using the Data Quality Objectives Process EPA QA / G-4,” no. February, 2006.

[3] Y. Jeon and Y. Kim, “Soil sampling plan design of key facilities for denuclearization based on data quality objective process,” *Nucl. Eng. Technol.*, vol. 54, no. 10, pp. 3788–3794, Oct. 2022, doi: 10.1016/j.net.2022.05.012.
[4] AI Hub, https://aihub.or.kr/
[5] Public data portal, https://www.data.go.kr/

ACKNOWLEDGEMENTS

This work was supported by the Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS), granted financial resources from the Nuclear Safety and Security Commission (NSSC), Republic of Korea. (No. 2106009) and Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.2020-0-01336, Artificial intelligence graduate school support(UNIST)).