# Metadata standardization in the Nuclear Energy Sector with Data Catalog Vocabulary

Young Ho Chae [a*], Yoonjoon Lee [b,c], Seo Ryong Koo [a]
[a]*Korea Atomic Energy Research Institute, 111 Daedeok-daero 989 beon-gil, Daejeon, 34057*
[b]*Korea Advanced Institute of Science and Technology, 291, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea*
[c]*SureDataLab Co. Ltd., 38, Expo-ro 446beon-gil, Yuseong-gu, Daejeon, Republic of Korea*
[*]*Corresponding author: yhchae@kaeri.re.kr*

***Keywords :*** standardization, data catalog vocabulary, artificial intelligence

## 1. Introduction

In the early stages, artificial intelligence(AI) based methodologies were confined to certain fields, constrained by issues related to the amount of data, calculation speed, and optimization algorithms. However, development from a various field enables AI application across diverse domains. In the field of nuclear, AI based applications are employed for the diagnosis of equipment, estimation of remaining useful lifetime(RUL), identification of accidents or abnormal situations [1-3].

The foundation for the development of domain-specific artificial intelligence (AI) is high-quality data. However, one of the major impediments to the efficiency of developing nuclear AI is issues related to data. The nuclear sector generates a vast amount of data across various domains such as plant operation, safety, and maintenance. Additionally, this data is managed differently across each power plant and simulation model. It is not standardized and often recorded in a format that is difficult for researchers to understand without the assistance of the data creators or responsible organizations, significantly hindering research efficiency.

Therefore, this study proposes a methodology for the metadata standardization of nuclear data using the Data Catalog Vocabulary (DCAT) as foundational research aimed at effectively designing and utilizing AI technology for the safe operation and maintenance of nuclear facilities. The standardization of metadata through DCAT is expected to offer the following advantages: 1) Provision of a standardized method for creating metadata, 2) Enhancement of data interoperability, 3) Improvement in data visibility and accessibility, 4) Increase in transparency and reliability, 5) Facilitation of data sharing among different entities, promoting collaborative and efficient use of data resources.

Continued research in this area is anticipated not only to accelerate the development and integration of nuclear AI technology but also to facilitate research across various domains within the nuclear field that require the utilization of diverse information, ultimately contributing significantly to the safety and efficiency of nuclear power plants.

## 2. Background

In this section, the characteristics of nuclear power plant (NPP) data, including simulation data, and the relevance of the DCAT for managing such data will be discussed.

### *2.1 Characteristics of NPP Data (Including Simulation Data)*

**Complexity and Diversity:** Nuclear power plants are intricate systems that generate a vast array of data from operations, maintenance, and safety monitoring activities. This data manifests in various forms, including sensor data, operational logs, states of different systems, and calculated values based on sensor data. The multifaceted characteristics of this data reflects the complex interactions and processes within a nuclear plant, necessitating sophisticated data management and analysis methodologies.

**High Reliability Required**: Data derived from nuclear power plants demands a high level of precision and reliability. Given the critical nature of nuclear energy production, any data utilized for the diagnosis of equipment, prediction of maintenance needs, or safety assessments must be accurate and trustworthy. This imperative underscores the need for rigorous data validation and quality control measures.

**Long-term Operation and Maintenance Data Management:** Nuclear power plants are designed for long-term operation, often spanning several decades. Throughout this duration, they accumulate extensive data sets that require effective management and analysis. The longevity of such data presents unique challenges in data storage, retrieval, and legacy system compatibility, emphasizing the importance of sustainable and scalable data management practices.

**Lack of Standardization Across Acronyms and Data Types Among Simulators:** A notable challenge within the nuclear sector is the absence of standardization across various acronyms and data types used in simulators and other data-generating sources. This inconsistency complicates data integration, comparison, and analysis processes, hindering collaborative research and operational efficiency.

The research scope encompasses all four characteristics. Nonetheless, in current stage, with the aim of evaluating the feasibility of applying DCAT, the focus of the research has predominantly been on data obtained from simulators.

### *2.2 Types of Metadata in the NPP simulation data*

This subsection describes representative metadata for NPP simulation data.

**Basic Information**

- Dataset Name: The unique name of the dataset.
- Description: A detailed explanation of the dataset's content, purpose, and scope.
- Creation Date: The date on which the dataset was created.
- Version: Information about the version of the dataset.
- Author/Organization: Details of the individual or organization that generated the dataset.

**Data Characteristics**

- Data Type: The type of simulation data (e.g., time-series data, spatial data).
- Variables and Parameters: A list and description of the variables and parameters used in the simulation.
- Units: The measurement units for each variable within the data.

**Simulation Settings**

- Simulation Model: The name and description of the simulation model used.
- Simulation Scenario: Description of specific conditions or scenarios for running the simulation.
- Start and End Time: The start and end times of the simulation.

**Simulation Results**

- Results Summary: An overview of the simulation results and key findings.
- Result Data Files: Location and format of files storing the simulation outcomes.

**Usage Conditions and Constraints**

- Copyright and Usage Rights: Legal constraints on the use of the dataset.
- Access Conditions: Conditions required to access the dataset.

**Related Documentation**

- Technical Documentation: Detailed documents on the simulation model, variables, algorithms, etc.
- References: Reference literature used in the research or design of the simulation.

The provision of such metadata supports researchers and engineers in the nuclear field to easily discover, understand, evaluate, and reuse simulation data. Enhancing data transparency and reusability, this approach facilitates knowledge sharing and collaboration through the simulation of various scenarios, thus advancing the field of nuclear engineering.

In addition to the basic forms of metadata for nuclear power plant (NPP) simulations, incorporating metadata commonly used in artificial neural network (ANN) training would significantly enhance data sharing and research. This enriched metadata set would include:

**Data Characteristics**

- Data Structure: Structure of the data.
- Variables and Parameters: Names, descriptions, and units of variables and parameters included in the dataset.

**Data Preprocessing and Cleaning Information**

- Preprocessing Information: Preprocessing steps applied to the data (e.g., normalization, noise removal, handling missing values).
- Preprocessing Parameters: Parameter values used in the preprocessing steps.

### *2.3 Introduction to Data Catalogue Vocabulary(DCAT)*

The DCAT is a World Wide Web Consortium (W3C) standard designed to facilitate interoperability between web-based data catalogs. Developed by the W3C, the primary goal of DCAT is to enable data publishers to increase the visibility and accessibility of their data to a wider audience. By providing a standardized model for describing datasets in a catalog, DCAT plays a crucial role in enhancing the discoverability and understanding of data available on the web. DCAT was developed with the intent to support the sharing, discovery, and use of data published on the web, regardless of the domain in which the data is published or the format in which it is stored. This standardization aims to make it easier for organizations and individuals to publish their data in a way that can be easily understood and consumed by others, promoting a more open and interconnected data ecosystem. Currently, DCAT is widely utilized across various sectors, including government, science, and education, to catalog datasets in a way that they can be easily found, accessed, and reused by interested parties. Government agencies around the world use DCAT to publish their open data, enabling citizens, businesses, and researchers to find valuable information for their needs. As of now, DCAT has been revised up to a third version [4]. The Fig.1 shows schematic diagram of DCAT V3 structure.

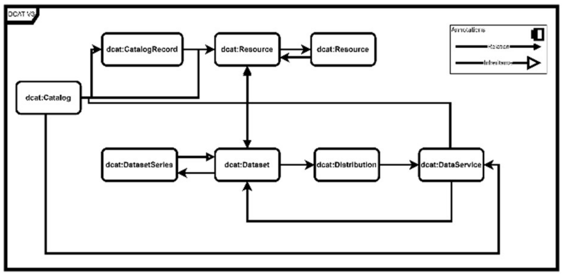


Fig. 1. Schematic diagram of DCAT V3.

## 3. DCAT-AP-KNPP Development

### *3.1 Requirements*

Incorporating the DCAT for managing metadata in nuclear power plant simulations necessitates adherence to several principle requirements to ensure effective utilization and interoperability. These requirements, crucial for the development and application of DCAT in this context, can be outlined as follows:

**Standardization of Metadata Descriptions**
Implementing DCAT requires the adoption of a standardized approach to describe datasets, ensuring consistent metadata across different simulations. This includes uniform descriptions of dataset names, variables, and parameters, facilitating easier data discovery and reuse.

**Comprehensive Data Documentation**
Detailed documentation of data, including its structure, preprocessing steps, and parameters, is essential. This documentation should cover all aspects of the data lifecycle, from collection through to processing and analysis, providing researchers with the necessary context for their work.

**Provenance and Quality Information**
It is vital to include metadata on the provenance and quality of the simulation data. This information should detail the source of the data, any transformations it has undergone, and assessments of its quality and reliability.

**Privacy and Security**
Given the sensitive nature of nuclear power plant simulations, the metadata schema must incorporate considerations for privacy and security. This includes mechanisms for controlling access to the data and ensuring that sensitive information is adequately protected.

### *3.2 Data Collections*

The types of datasets utilized in this study include data acquired from three specific nuclear power simulators and one thermal-hydraulic analysis code: (1) Compact Nuclear Simulator (CNS), (2) PCTRAN Simulator, and (3) 3-Key Master Simulator, alongside (4) MARS(Multi-dimensional Analysis of Reactor Safety) analysis data specifically  for an OPR-1000 power plant.

The CNS mathematical model simulates a three-loop Westinghouse Pressurized Water Reactor (PWR), with an output capacity of 993 MWe. The PCTRAN Simulator models a two-loop Korea Hydro & Nuclear Power (KHNP) OPR-1000 PWR, capable of generating 1000 MWe. The 3-Key Master Simulator's mathematical model represents a two-loop KHNP APR-1400 PWR, with a higher capacity of 1400 MWe. Lastly, the TH Code Analysis Data, specifically using the MARS code, focuses on modeling the OPR-1000 power plant.

### *3.3 Applying DCAT to NPP Simulation Data*

The structure of DCAT-KNPP can be broadly divided into two main parts: the Core part and the Extension part. The Core part contains fundamental elements that are shared across various types of data and form the basic components of a data catalog. Structurally, the Core part of DCAT-KNPP is primarily based on DCAT Version 3. To facilitate one of the key functionalities of DCAT-KNPP, which is search capability, it incorporates an additional Simple Knowledge Organization System (SKOS) class. In the diagram below, classes represented in white boxes are the basic classes from DCAT Version 3, and those in yellow boxes signify the SKOS classes added to enhance search functionality (Fig.2, Top).

The Extension part of DCAT-KNPP is designed to be flexible, allowing for the addition of elements tailored to the specific characteristics of the data being cataloged or the type of data one wishes to metadata. For instance, if a description of simulation software is considered a critical piece of information for the catalog, a new 'knpp software' class can be established, under which various properties can be detailed. Fundamentally, all extensions inherit from the core part's Resource class and are linked to the Catalog class, enabling their inclusion in the data catalog. The relationship between the core and extension parts is illustrated in the figure below, showcasing how the structure facilitates the seamless integration of additional information into the data catalog while maintaining a robust foundation for data organization and searchability (Fig.2, Bottom).

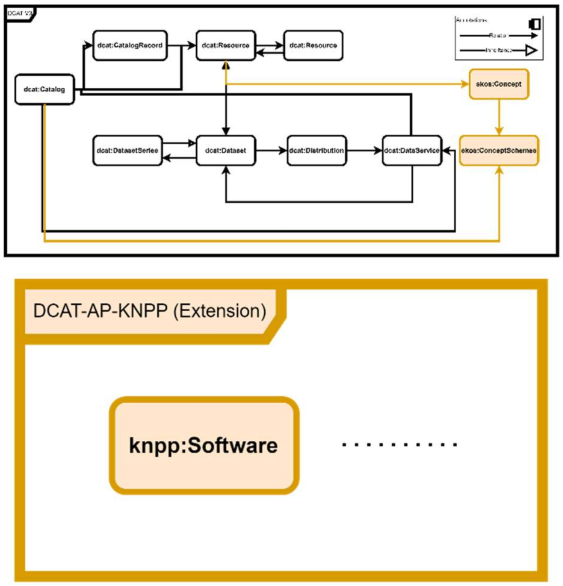


Fig. 2. DCAT-AP-KNPP Core, Extension

## 4. Pilot Experiment

In this section, a simplified scenario is showcased, based on a DCAT environment that comprises metadata for 12 datasets, organized into a catalog. The structure of the NPP metadata is based on the RDF(Resource Description Framework) format. By adopting a logical approach to describing metadata using RDF (Resource Description Framework), to demonstrate the potential for efficient data management and retrieval in such complex domains.

A brief description of a group of files reveals that it is divided into 5 datasets. For example, the file named "A-TH-15-6-5-SBLOCA-CL-CHA-0.9INCH.csv" is part of a dataset group that simulates a nuclear power plant using the OPR1000 reactor model. This group is further categorized into nodalization, node-description, initial conditions, event time table, and output dataset. The schematic diagram of the structure of the dataset is shown in Fig.3.

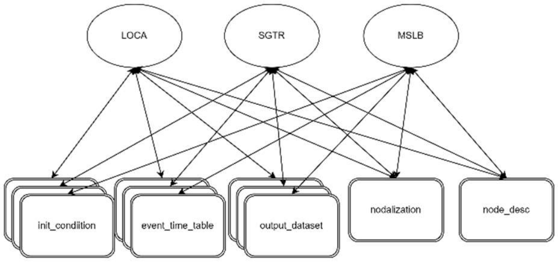


Fig. 3. Structure of the example dataset

Furthermore, the Simple Knowledge Organization System (SKOS) [5] has been applied to enable semantic search within these catalogs, developing a taxonomy that classifies the catalogs in a more meaningful way. The example SKOS for NPP simulation data is shown in Fig. 4.

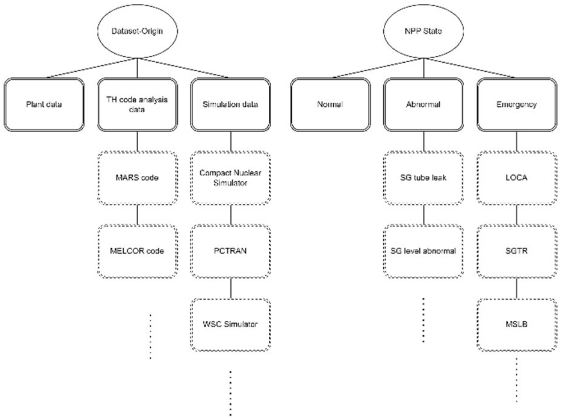
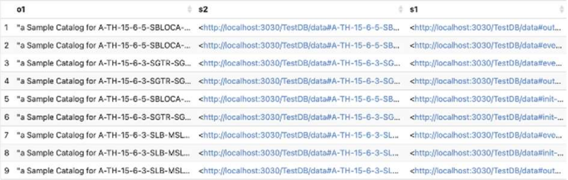


Fig. 4. SKOS for NPP simulation data

Through the integration of metadata structuring and the application of SKOS, a metadata-based search system was developed. As part of an experiment, an attempt was made to retrieve a list of datasets that simulate an 'emergency' state in Korean ('비상') through a query. The results of this experiment are showcased in Figure 5. With this system, data searches based on metadata can be successfully conducted. This example not only underscores the practicality of the approach but also highlights the flexibility and power of combining DCAT-based metadata with RDF for advanced data discovery and management in critical infrastructure sectors like nuclear energy.



| | o1 | s2 | s1 |
|---|---|---|---|
| 1 | "a Sample Catalog for A-TH-15-6-5-SBLOCA-... | <http://localhost:3030/TestDB/data#A-TH-15-6-5-SB... | <http://localhost:3030/TestDB/data#out... |
| 2 | "a Sample Catalog for A-TH-15-6-5-SBLOCA-... | <http://localhost:3030/TestDB/data#A-TH-15-6-5-SB... | <http://localhost:3030/TestDB/data#eve... |
| 3 | "a Sample Catalog for A-TH-15-6-3-SGTR-SG... | <http://localhost:3030/TestDB/data#A-TH-15-6-3-SG... | <http://localhost:3030/TestDB/data#eve... |
| 4 | "a Sample Catalog for A-TH-15-6-3-SGTR-SG... | <http://localhost:3030/TestDB/data#A-TH-15-6-3-SG... | <http://localhost:3030/TestDB/data#out... |
| 5 | "a Sample Catalog for A-TH-15-6-5-SBLOCA-... | <http://localhost:3030/TestDB/data#A-TH-15-6-5-SB... | <http://localhost:3030/TestDB/data#init-... |
| 6 | "a Sample Catalog for A-TH-15-6-3-SGTR-SG... | <http://localhost:3030/TestDB/data#A-TH-15-6-3-SG... | <http://localhost:3030/TestDB/data#init-... |
| 7 | "a Sample Catalog for A-TH-15-6-3-SLB-MSL... | <http://localhost:3030/TestDB/data#A-TH-15-6-3-SL... | <http://localhost:3030/TestDB/data#eve... |
| 8 | "a Sample Catalog for A-TH-15-6-3-SLB-MSL... | <http://localhost:3030/TestDB/data#A-TH-15-6-3-SL... | <http://localhost:3030/TestDB/data#init-... |
| 9 | "a Sample Catalog for A-TH-15-6-3-SLB-MSL... | <http://localhost:3030/TestDB/data#A-TH-15-6-3-SL... | <http://localhost:3030/TestDB/data#out... |

Fig. 5. Search results

## 5. Limitations and Future Work

The contribution of data creators is crucial in the context of metadata. It is because metadata must be appropriately tagged by the creators to be utilized in searches. However, this requirement can impose a significant burden on the data creators. To address this issue, it is anticipated that the widespread adoption of the metadata system would be facilitated if a system were implemented that could automatically generate most of the content upon receiving a dataset sample. The research conducted so far has been limited to applying this approach to a restricted set of data to ascertain its feasibility. Future research will aim to extend this approach to a broader array of dataset types. Additionally, plans are in place to design an automatic metadata generator, as mentioned in the limitations section.

## Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144150)

## REFERENCES

[1] Lee, Gyumin, Seung Jun Lee, and Changyong Lee. "A convolutional neural network model for abnormality diagnosis in a nuclear power plant." Applied Soft Computing 99 (2021): 106874.
[2] Shin, Ji Hyeon, et al. "An interpretable convolutional neural network for nuclear power plant abnormal events." Applied Soft Computing 132 (2023): 109792.
[3] Chae, Young Ho, et al. "Graph neural network based multiple accident diagnosis in nuclear power plants: Data optimization to represent the system configuration." Nuclear Engineering and Technology 54.8 (2022): 2859-2870.
[4] https://www.w3.org/TR/vocab-dcat-3/
[5] https://www.w3.org/2004/02/skos/