기본연구보고서 2025-8

인공지능을 활용한 건축물 데이터 품질 고도화 방향 연구

Enhancing Building Data Quality Using Artificial Intelligence

안의순 Ahn, Euisoon
허한결 Heo, Hankyul
남기천 Nam, Kicheon
강범준 Kang, Bumjoon
이선재 Lee, Sunjae
박동준 Park, Dongjoon

Summary

Enhancing Building Data Quality Using Artificial Intelligence

Ahn, Euisoon Heo, Hankyul Nam, Kicheon Kang, Bumjoon Lee, Sunjae Park, Dongjoon

Introduction

Buildings are fundamental spatial units that form the basis of policymaking, so having an accurate picture of the nation's building stock is a critical task. However, data related to buildings produced and managed by the government are limited in both quality and interoperability. The building register—the representative administrative dataset for buildings—is an official ledger that records current building information and serves as the core foundation for inter-ministerial data integration on buildings. Nevertheless, previous studies have repeatedly pointed out issues of errors, omissions, and inconsistencies in these datasets.

A complete on-site survey of all buildings would be the only way to fully ascertain the actual building status, but such an approach would be prohibitively costly and impractical. Therefore, this study aimed to verify building data errors and identify data quality issues such as omissions, outliers, and inconsistencies by applying artificial intelligence (AI) methodologies that minimize human intervention. Furthermore, to strengthen the technical and institutional foundation for the open access of building data, the study explored data linkage methods based on "Building ID"—a unique building-level identifier introduced in the "Seumter," or the e-AIS, the electronic Architectural Information System.

Accordingly, this study explored AI-based strategies to improve the quality of building data and proposed practical applications for quality enhancement. It also suggested measures to distribute quality-enhanced datasets and improve the quality of newly generated data during the building permit process.

Directions for Enhancing Building Data Quality

■ Current state of legislation and regulations regarding building data

Under the Building Act, building permit processes must be handled via an electronic information processing system (Article 32), and permit authorities must maintain and update the Building Register to ensure accurate building information (Article 38). Accordingly, the Seumter system has been in operation.

Also under the Act on Promotion of the Provision and Use of Public Data, public institutions are required to provide the data they manage so that they may be utilized by the public (Article 3). Unprocessed administrative building data managed through Seumter are provided accordingly via the Building Hub platform. However, this data, along with the original register, lacks continuity and consistency.

In addition, the primary key (PK) structure has changed since the transition from the previous Open Building Data System to the new Building Hub in early 2025, making it difficult to link data across the time series.

■ Quality and Utilization of Building Data

The Ministry of Land, Infrastructure and Transport (MOLIT) and local governments have been continuously refining the building register. In 2022, Seumter defined 86 rules for inspecting the building register along with building/housing permit data, and reviewed their error rates. Out of 620 million total records, about 8.55 million contained errors, corresponding to an overall error rate of 1.37%. However, some rules showed much higher error rates between 10% and 30%.

Building data are widely used for statistical and research purposes. Current construction statistics are automatically generated through the administrative system. In these datasets, key variables include district codes, floor area, ownership type, usage codes, and floor counts, while permit-related datasets use district codes, usage and structure codes, floor area, type of permit, number of buildings, and key approval dates.

A review of prior research revealed common data errors, including invalid main-use codes, missing building names or main-building counts, and unrealistic outliers in total floor area.

■ Directions for Building Data Quality Enhancement

Based on the review of current quality and utilization, this study proposes strategies to enhance data on building area and usage, as well as linkage quality between datasets.

The building area data is a numerical variable central to statistics and research. Since the building register shows the highest error rate in rules on area data, it was prioritized for review. Four rules from the Seumter inspection (site area, building area, building coverage ratio, and floor area ratio) were verified again as of now and analyzed by region and year of completion. In addition, new verification rules were developed to supplement the existing framework.

To go beyond rule-based checks, AI-based anomaly detection techniques such as Isolation Forest and One-Class SVM were employed. These were applied complementarily to detect potential errors that cannot be captured through the rule-based method.

The building usage data were also analyzed, as they are widely used in both statistics and research. To ensure consistency between free-text "other use" entries and coded "main use" classifications, a Naive Bayes classifier was trained to predict main use codes from textual descriptions.

Finally, the study examined data linkage quality using Building ID by connecting building registers with building and housing permit datasets, and analyzing linkage rates, data consistency, and matching success.

Pilot Application of Building Data Quality Enhancement

■ Building Register Area Data

A comprehensive diagnostic process combining existing rule-based validation, newly developed rules, and AI-based anomaly detection revealed an error rate ranging from 0.13% to 5.01%.

New validation rules were added for cases where the floor area used for floor area ratio (FAR) calculation exceeded the total floor area (0.26%), and where the total floor area exceeded the derived upper threshold (0.86%).

For anomaly detection, dimensionless indicators were created—such as the ratio of building area to site area, the ratio of FAR floor area to total floor area, and floor occupancy ratios. Results showed temporal stability but different regional variation pattern from the existing and new rules: higher anomaly rates were found in Seoul and Busan, suggesting that the AI approach successfully detected previously unidentifiable error patterns.

Building Register Usage Data

The usage data were systematically analyzed for quality issues. Frequency analysis revealed mismatches between frequently occurring words and official use codes, particularly in mixed-use buildings. Using a Naive Bayes classifier, prediction accuracy reached 84.78% at the building level and 78.97% at the floor level. However, class imbalance and multiple-use entries led to misclassification, highlighting the need for layer-specific verification across all floors.

Linking between Building Register and Permit Data

A Building ID–based linked dataset was established to assess linkage quality between the building register and building/housing permit data. Among 55,784 buildings approved for occupancy in 2024, Building ID assignment rates were 97.0% for the building register, 99.4% for building permits, and 99.4% for housing permits.

Linkage success rates were 93.6% for building permits and 34.9% for housing permits. Among successfully linked records, area consistency exceeded 90% for building permits but ranged from 67–73% for housing permits; usage data consistency exceeded 95% for both datasets.

A shift toward a tailored maintenance framework that accounts for error patterns in building data

The pilot application confirmed that building data errors are not random but structured by temporal, regional, and institutional factors. Hence, uniform periodic maintenance alone is insufficient. A context-aware and customized maintenance framework integrating AI-based anomaly detection with rule-based validation is required.

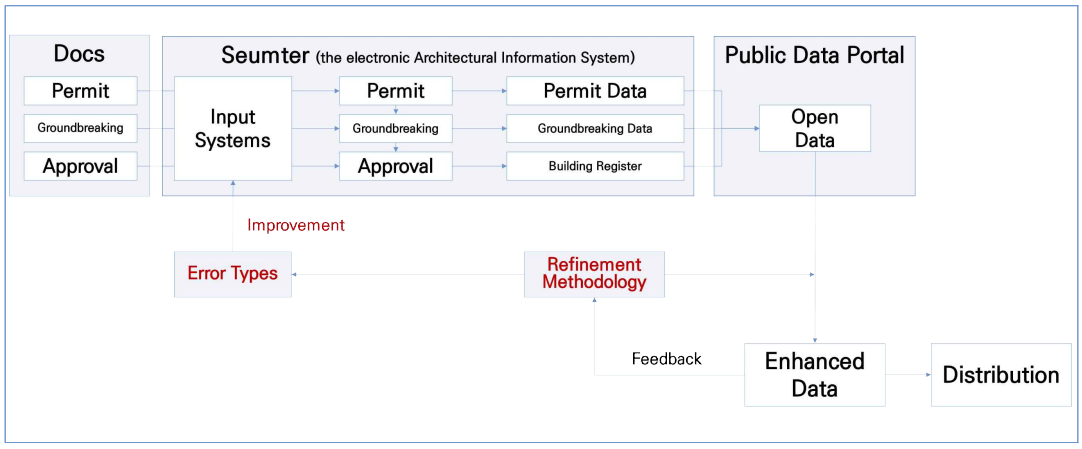
Furthermore, Building ID needs stricter management of linkage quality, ensuring the uniqueness and reliability of Building IDs as universal linkage keys across datasets.

Proposals for Building Data Quality Improvement

Independent Distribution of Enhanced Building Data

Given the administrative difficulty of directly amending official registers, this study proposes maintaining enhanced data externally from Seumter and distributing them as open-source datasets. Also the methodologies themselves—(1) rule-based error detection, (2) AI-based anomaly detection, and (3) cross-verification using Building IDs—should be openly published as reproducible documentation and code.

A phased release through the Public Data Portal is recommended, incorporating user feedback and version control to ensure continuous improvement and sustainability of the enhancement.



Proposals for Building Data Quality Improvement
Source: Written by the research team

Improving the Quality of New Data

For newly constructed buildings, validation should be integrated into the building and floor summary input systems during the permit process. Area values, currently entered manually, are prone to errors; therefore, automatic calculations and alert mechanisms should be introduced when discrepancies arise. For usage-related fields, inconsistencies or typos between free-text entries ("other use") and classified codes should trigger warnings for user correction.

At the approval for use stage, Building IDs should be actively used as linkage keys to verify data consistency between permit data and the building register data. When inconsistencies occur, the system should alert users, providing an opportunity to correct the data while maintaining the authority of the permitting agency.

Keywords
Building Register, Building ID, Machine Learning, Public Data, Open Source