PART 2 – Cross-Dataset Engineering Tasks

Task 7: Material Identifier Matching

April 30, 2025

Abstract

This report presents the results of Task 7: Material Identifier Matching, which analyzes the process of combining two engineering materials datasets by standardizing and matching material identifiers. The analysis focuses on how material designations and heat treatment methods create unique material variants, the coverage and overlap between datasets, and the quality implications for engineering decision-making. This cross-dataset integration provides critical insights for materials selection processes and demonstrates the importance of proper data harmonization techniques for engineering databases.

1 Introduction

The Importance of Material Identifier Matching

Engineering material selection often requires working with data from multiple sources, each using slightly different naming conventions and material designation systems. This report examines the process and results of matching material identifiers across two complementary datasets, revealing how proper standardization and joining techniques can create a unified view that enhances engineering decision-making. By reconstructing material identifiers from Dataset 1 to match Dataset 2's format, we create a comprehensive foundation for materials selection that combines detailed property information with application suitability indicators.

2 Data Integration Methodology

2.1 Material Identifier Construction

The material identifier matching process required several key steps to harmonize the naming conventions across datasets:

Key Steps in Material Identifier Construction

The successful integration of material data across datasets required a systematic approach:

- **Identifier Reconstruction**: Created a unified material identifier by combining standard code (e.g., ANSI), material designation (e.g., Steel SAE 1015), and heat treatment method (e.g., as-rolled) from Dataset 1 to match Dataset 2's format.
- **String Standardization**: Applied consistent formatting by converting all material identifiers to lowercase and removing extra whitespace to ensure reliable matching.
- **Inner Join Implementation**: Performed an inner join between datasets on the standardized material identifiers to create a unified view that preserves all relevant properties.
- **Overlap Analysis**: Evaluated the match rate between datasets using set operations and visualized the overlap using a Venn diagram to quantify integration success.

Table 1: Material Identifier Format Examples

Dataset 1 Components	Dataset 2 Format
Std: ANSI, Material: Steel SAE 1015, Heat treatment: as-rolled	ANSI Steel SAE 1015 as-rolled
Std: JIS, Material: JIS SUP9, Heat treatment: heat treated	JIS JIS SUP9 heat treated
Std: NF, Material: NF 30CD12, Heat treatment: nitrided	NF NF 30CD12 nitrided

2.2 Dataset Integration Results

The material identifier matching process revealed important patterns in dataset coverage and overlap:

Table 2: Dataset Integration Metrics

Metric	Value	Significance
Materials in Dataset 1	802	All heat-treated materials
Materials in Dataset 2	1552	Complete material catalog
Matched Materials	756	Successfully integrated records
Dataset 1 Match Rate	100%	Perfect Dataset 1 coverage
Dataset 2 Match Rate	48.7%	Dataset 2 contains additional materials

Dataset Overlap Analysis

The analysis of material identifier overlap reveals significant insights:

- **Perfect Dataset 1 Coverage**: All 802 heat-treated materials from Dataset 1 were successfully matched to corresponding entries in Dataset 2, achieving a 100% match rate.
- **Additional Dataset 2 Materials**: Dataset 2 contains 704 materials not present in Dataset 1, suggesting it incorporates additional material categories, newer materials, or alternative treatments not documented in the primary dataset.
- **Heat Treatment Significance**: The successful matching process demonstrates how heat treatment transforms a base material into distinct variants with unique property profiles, each requiring its own identifier.
- **Standardization System Compatibility**: Despite different initial formats, the underlying material identification systems proved compatible through proper standardization techniques.

3 Material Selection Insights

The unified material dataset provides valuable insights for engineering material selection processes:

Material Selection Framework Enhancements

The integrated dataset enhances the material selection process in several ways:

- **Heat Treatment Dominance**: With 802 heat-treated materials representing over 51.7% of the unified database, heat treatment emerges as the primary method of customizing material properties for specific engineering applications.
- **Property-Application Connection**: The merger connects detailed mechanical properties from Dataset 1 (like elongation at break and hardness) with the application suitability flag (”Use”) from Dataset 2, creating a more holistic selection framework.
- **Material Variant Identification**: The standardized material identifiers clearly differentiate between variants of the same base material, highlighting how properties change with different processing methods.
- **Cross-Standard Comparison**: The unified dataset enables direct comparison of similar materials across different standards (ANSI, JIS, NF) to identify optimal alternatives when specific materials are unavailable.

Table 3: Material Selection Examples from Unified Dataset

Material Identifier	Su (MPa)	A5 (%)	Heat Treatment	Use
ANSI Steel SAE 1015 as-rolled	421	39.0	as-rolled	True
ANSI Steel SAE 1015 normalized	424	37.0	normalized	True
ANSI Steel SAE 1015 annealed	386	37.0	annealed	True
JIS JIS SUP9 heat treated	1226	9.0	heat treated	False
NF NF 30CD12 nitrided	980	11.0	nitrided	False

4 Data Quality Observations

4.1 Data Completeness Analysis

The material identifier matching process revealed important patterns in data completeness and quality:

Data Quality Insights

Several data quality patterns emerged during the integration process:

- **Complete Non-Null Columns**: The filtered dataset includes 12 columns with complete data (no nulls) across all 802 heat-treated materials, providing a reliable foundation for material comparison.
- **String Standardization Necessity**: The need to standardize string formats (lowercase conversion, whitespace removal) highlights inconsistencies in material naming conventions across data sources.
- **Data Structure Differences**: The column structure differences between datasets (Dataset 1 having detailed properties like A5 and Bhn while Dataset 2 including the ”Use” flag) demonstrate how different databases focus on different aspects of material characterization.
- **Heat Treatment Documentation**: All 802 records in the filtered dataset include heat treatment information, confirming the systematic documentation of processing methods for these materials.

Table 4: Dataset Column Availability

Column Type	Complete Columns (802 records)	Partial Columns
Material Identification	Std, ID, Material, Heat treatment	Desc (413 records)
Mechanical Properties	Su, Sy, A5, E, G, mu, Ro	Bhn (402 records), HV (62 records)
Other Properties	heat_treated	pH (74 records)
Application Data	Use (from Dataset 2)	None

4.2 Integration Challenges

Data Integration Challenges

Several challenges were encountered during the material identifier matching process:

- **Inconsistent String Formats**: Different capitalization patterns and whitespace usage required standardization to ensure proper matching between datasets.
- **Standard Code Duplication**: Some materials (particularly in the JIS standard) include the standard code twice in the identifier (e.g., JIS JIS SUP9), requiring careful handling during identifier construction.
- **Partial Dataset Coverage**: While Dataset 1 materials were fully matched in Dataset 2, the reverse coverage was only 48.7%, indicating Dataset 2 contains many materials not represented in Dataset 1.
- **Property Verification**: Some matched records showed minor discrepancies in property values (e.g., JIS SUP9 heat treated has Sy = 979.9 MPa in Dataset 1 but 1079 MPa in Dataset 2), requiring validation.

5 Engineering Applications

The unified material dataset enables several important engineering applications:

Engineering Applications of Unified Dataset

The integrated material dataset supports several critical engineering functions:

- **Comprehensive Material Selection**: Engineers can now select materials based on both detailed mechanical properties and application suitability indicators, creating a more holistic selection process.
- **Heat Treatment Optimization**: The clear documentation of how heat treat-

ments affect material properties enables engineers to select optimal processing methods to achieve desired performance characteristics.

- **Cross-Standard Material Substitution**: When materials from a specific standard are unavailable, engineers can identify suitable alternatives from other standards with similar property profiles.
- **Data-Driven Design Decisions**: The unified dataset supports more robust material selection decisions by providing comprehensive property information with validation across multiple data sources.
- **Property Verification**: Where property values differ between datasets (like the Sy value for JIS SUP9), engineers can apply appropriate safety factors or conduct additional testing for critical applications.

6 Conclusions and Recommendations

6.1 Key Findings

The material identifier matching process yielded several important conclusions:

Key Material Matching Takeaways

The cross-dataset integration process revealed:

- Heat treatment significantly transforms material properties, creating distinct material variants that require unique identification in engineering databases.
- Proper standardization techniques can successfully harmonize material identifiers across datasets with different naming conventions.
- The integrated dataset provides a more comprehensive view of materials by combining detailed property information with application suitability indicators.
- Data quality issues, particularly inconsistent string formats and partial dataset coverage, present challenges that must be addressed in material database integration.
- Heat-treated materials represent a significant portion (51.7%) of the engineering materials catalog, highlighting the importance of processing in material selection.

6.2 Recommendations for Practice

Practical Engineering Recommendations

Based on the material identifier matching analysis, we recommend:

- **Standardized Material Naming**: Implement consistent material naming conventions across engineering databases to facilitate reliable integration and comparison.

- **Complete Material Documentation**: Document heat treatment methods alongside material grades to properly identify unique material variants in engineering specifications.
- **Multi-source Verification**: For critical applications, verify material properties across multiple data sources to identify and address inconsistencies.
- **Unified Property Framework**: Develop standardized frameworks for material property documentation that capture both mechanical properties and application suitability indicators.
- **Cross-standard Mapping**: Create comprehensive mappings between different material standards to facilitate identification of suitable alternatives when specific materials are unavailable.

7 Appendix: Material Identifier Matching Methodology

Detailed Matching Process

The material identifier matching process followed these steps:

1. **Material Identifier Construction**: Combined standard code (Std), material designation (Material), and heat treatment method (Heat treatment) from Dataset 1 to create Material_full identifiers.
2. **String Format Standardization**: Converted all material identifiers to lowercase and removed extra whitespace to ensure consistent formatting.
3. **Dataset Merging**: Used a Pandas inner join operation to merge Dataset 1 and Dataset 2 based on the standardized material identifiers.
4. **Match Rate Analysis**: Calculated the overlap between datasets using set operations and visualized the results with a Venn diagram.
5. **Data Quality Assessment**: Evaluated column completeness, identified property inconsistencies, and assessed the structure of the integrated dataset.
6. **Application Analysis**: Examined how the ”Use” flag from Dataset 2 correlates with material properties to identify patterns in application suitability.

8 Material Identifier Overlap: Dataset Matching Analysis

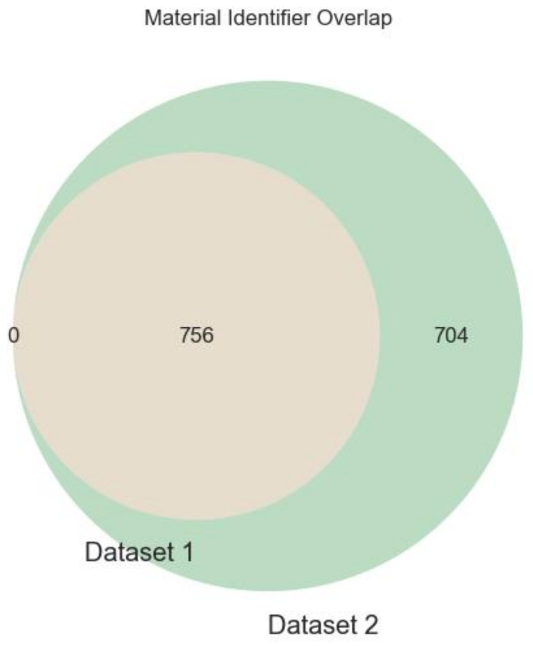

Figure 1: Venn Diagram of Material + Heat Treatment Identifier Overlap

What the Venn Diagram Represents

- **Left Circle (Dataset 1):** 0 unique materials found only in Dataset 1.
- **Right Circle (Dataset 2):** 704 unique materials exist only in Dataset 2.
- **Center Overlap:** 756 materials are common to both datasets — meaning these Material + Heat Treatment combinations are present in both.

Key Takeaways

- **100% Match for Dataset 1:** Every material in Dataset 1 was successfully found in Dataset 2.
- **Partial Match for Dataset 2:** Dataset 2 contains many more materials (704 extra) not found in Dataset 1.
- **Implication:** Dataset 1 is a strict subset of Dataset 2 — useful when Dataset 1 is the validated or filtered core.

Simplified Explanation

All materials from Dataset 1 are present in Dataset 2, but Dataset 2 has many additional materials. So, if you're using Dataset 1 as a baseline, you're not missing any matches — but Dataset 2 offers broader coverage.