# Incorporation of the Uniformat II classification system

## Introduction

The primary source for Uniformat II data is a comprehensive but unstructured PDF guide. The objective of this workstream was to design and implement an automated pipeline to transform the static Uniformat II guide into a structured, enriched, and queryable SQLite database. This database serves as a foundational asset for the next project phase: mapping historical project costs to Uniformat II codes.

## Methodology

The Uniformat II guide was manually parsed to extract the proposed classification system in the following columns:

* Type, e.g. ‘Building’
* Level 1 Code, e.g. ‘A’
* Level 1 Name, e.g. ‘Substructure’
* Level 2 Code, e.g. ‘A10’
* Level 2 Name, e.g. ‘Foundations’
* Level 3 Code, e.g. ‘A1010’
* Level 3 Name, e.g. ‘Standard Foundations’
* Level 4 Code, e.g. ‘A1011’
* Level 4 Name, e.g. ‘Wall Foundations’

The extraction if inclusion/exclusion data and description enrichment were handled with an automated pipeline divided into three main steps: database initialization, AI-driven data extraction, and AI-powered content enrichment. The process was executed through a Jupyter Notebook that orchestrated several modular Python scripts, ensuring a repeatable and transparent workflow.

The process for converting the Uniformat II guide into the database is shown below.

A diagram of a company

AI-generated content may be incorrect.

### Database Initialization and Baseline Loading

First, a local SQLite database named uniformat.db was created. The schema was defined using db\_operations.py to include three core tables:

* uniformat\_codes: The main table to hold the hierarchy of Uniformat codes and their names.
* uniformat\_inclusions: A table to store specific items included under a given code.
* uniformat\_exclusions: A table to store items explicitly excluded from a code's scope.

The uniformat\_codes table was then populated with baseline data from a uniformat-ii-codes.csv file, which contained the code and name hierarchy from Level 1 to Level 4. This provided the structural skeleton for our database.

### Structured Data Extraction from PDF via LLM

The central challenge was converting the unstructured "Includes" and "Excludes" lists from the Uniformat guide PDF into structured data. The process began by extracting raw text from the relevant section of the document, specifically pages 61 through 83 (which contained the information about inclusions and exclusions). This text was then provided to the Google Gemini API, which was instructed to act as a construction classification expert. A key step was constraining the model to return the parsed information—each element's code, name, inclusions, and exclusions—as a strictly formatted JSON object. This machine-readable output was then used to populate the corresponding uniformat\_inclusions and uniformat\_exclusions tables in the database.

With the foundational data structured, the next phase focused on enrichment. All the information for the 79 unique Level 3 elements was aggregated from the database, combining the codes, names, and the newly added inclusions and exclusions. This aggregated context was sent back to the Gemini API in a second, distinct step. Here, the model's task shifted from extraction to synthesis. The model was prompted to generate a comprehensive, professional description for each element based on the full context provided. To ensure efficient and stable processing, these requests were sent in batches of five. Finally, these newly generated descriptions were written back into the main uniformat\_codes table, completing the data enrichment pipeline.

## Results and Verification

The pipeline successfully produced a self-contained SQLite database (uniformat.db) containing fully enriched data for all 79 Level 3 Uniformat II elements. The final database now includes:

* A complete hierarchy of Uniformat II codes and names.
* Populated description fields for every Level 3 code.
* Associated inclusions and exclusions tables, providing essential details for accurate classification.

For example, the final generated description for code A1010, Standard Foundations, is:

*“This element encompasses the complete scope of standard shallow foundation systems designed to transfer building loads to the supporting soil. It includes the construction of isolated column footings and continuous wall foundations, along with their associated bases. Integral to this element are foundation walls that extend from the footing up to the level of the top of the slab on grade, providing essential support and enclosure. The scope also covers the formation of concrete pile caps, which serve to distribute loads from columns or walls to a group of piles. Essential earthwork activities such as backfilling around the foundation elements and ensuring proper compaction are included. Furthermore, this category incorporates critical protective and functional components like perimeter insulation to mitigate heat loss and perimeter drainage systems designed to manage subsurface water around the foundation. The installation of anchor plates, crucial for securing structural elements to the foundation, is also part of this element, as well as any dewatering operations specifically required during the construction of these standard foundation components.”*

The resulting database is now ready to be used to enrich the construction project data. This will likely involve an embedding-based approach to semantically map the existing cost items to the new Uniformat II descriptions, creating a richer dataset for the predictive models.