IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 005, LIR419
DOI: https://doi.org/10.52731/lir.v005.419

Applying Text Generation AI to Assist Categorical Database Construction from Institutional Regulations

Tsunenori Inakura [*], Shotaro Imai [†], Kunihiko Takamatsu [*], Sayaka Matsumoto [*], Masao Mori [*]

Abstract

We demonstrated that a text generation AI (ChatGPT) can support the construction of categorical databases from institutional documents. By extracting concepts, formulating functional relationships, and verifying the semantic validity of composite morphisms consistency through natural language, this approach supports the creation of database schemas that accurately reflect the meaning of the original documents. It also makes it easier for non-experts to take part in schema building, providing a clear way to transform written regulations into structured data. This semi-automated method not only reduces manual workload but also improves clarity and maintainability of database structures. Our findings highlight the potential of language models to bridge formal data modeling and natural language logic in educational and administrative domains.

Keywords: Database, Categorical database, Text generation AI

1 Introduction

Institutional Research (IR) activities in universities aim to support decision-making by collecting and analyzing data related to education, research, and university operations. IR relies on data warehouses (DWs) that collect data from various internal systems, offering an integrated foundation for analysis. However, maintaining data consistency and usability across different departments remains a persistent challenge.

Many databases (DBs) within universities are individually built to serve specific operational needs, without prioritizing broader analytical integration. As these systems are updated independently over time to meet the needs of day-to-day operations, it becomes increasingly difficult to preserve the logical structure of data references. Even when a DW is carefully designed at the outset, ongoing independent modifications to individual systems often disrupt the consistency between DBs and DW, undermining the reliability of data for IR purposes. In addition, many university systems are operated by administrative staff rather than database specialists, making it even more difficult to maintain database integrity in the long term.

To design and manage DBs and DWs effectively, expertise is required in translating operational requirements into systematic data structures, as well as in maintaining the logical consistency of entity-relationship diagrams (ERDs). However, this level of expertise is not available across all departments, and the daily pressures of university administration often led to ad hoc modifications that compromise the consistency of the overall data structure.

[*] Institutional Research Section, Office of Institute Strategy, Institute of Science Tokyo, Tokyo, Japan
[†] Center for Information Infrastructure, Institute of Science Tokyo, Tokyo, Japan

In response to these challenges, categorical databases [1,2] have been proposed as a theoretically robust solution. Category theory, a branch of mathematics, provides a framework for describing objects and relationships in a structured and consistent way. A categorical database applies these principles to database design, ensuring that as long as modifications respect the defined structures, data consistency can be preserved even as systems evolve.

One particularly accessible way to design categorical databases is through the use of ologs (ontology logs) [3]. Ologs allow the structure of a database to be described using natural language, making it easier for non-specialists to participate in schema design. In an olog, concepts are represented as objects, and relationships between them are expressed as natural language sentences. This approach enables non-experts to validate whether the intended data references are logically correct based on the semantics of the domain, without needing deep knowledge of formal category theory.

By representing database structures in olog form, it becomes possible to generate conceptual diagrams that can be directly mapped to ER diagrams and database schemas. This natural language-based modeling approach not only improves the intuitiveness and accessibility of database design but also supports the construction of highly consistent data models that can evolve flexibly over time.

We explored the potential of using a text generation AI (ChatGPT) to support the semi-automated construction of categorical database schemas from natural-language institutional documents. By guiding the model through concept extraction, relationship definition, and structure verification processes grounded in category-theoretic principles, we aimed to make database modeling easier by reducing the amount of technical knowledge and complex reasoning usually required. At the same time, we tried to keep the accuracy and consistency needed for categorical databases. Our approach centered on leveraging dialogue-based natural-language interaction to make schema construction more accessible, particularly for non-specialists.

2 Categorical Database

A categorical database is a way of organizing data that focuses not only on individual data items, but also on the relationships between them. Based on ideas from category theory, it models the structure of concepts and their connections in a rigorous and consistent manner.

In a categorical database, data is structured as a collection of objects (concepts) and morphisms (relationships) between them. Each morphism must be functional, meaning that every instance of the source concept corresponds to exactly one instance of the target concept. This requirement ensures that relationships are well-defined and easy to trace.

An olog (ontology log) provides a practical and intuitive framework for describing a categorical database. Figure 1 shows an example of olog. In an olog, each relationship between concepts is expressed as a simple, natural language sentence. This design allows even non-specialists to understand and verify the structure without needing formal mathematical knowledge.

An important feature of categorical databases is composability. When two morphisms are connected in sequence — for example, from A to B and then from B to C — it must be possible to compose them into a single, meaningful relationship from A to C. Checking that these composite morphisms are semantically valid ensures that the overall structure remains consistent, even when data is referenced across multiple steps.

By using ologs and validating the composability of relationships, it becomes possible to maintain data integrity across complex structures. This approach is especially valuable when constructing

databases that integrate information from different systems over time, such as in institutional research environments.

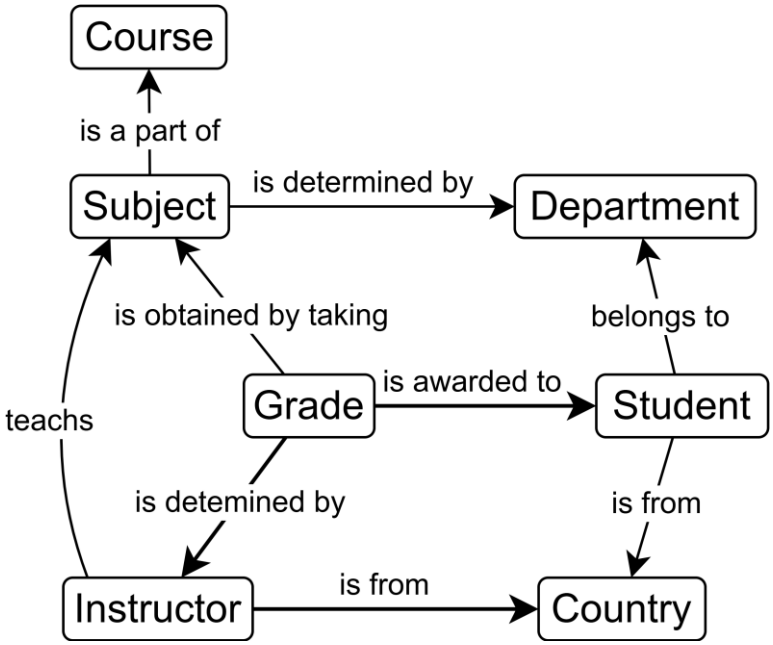

Figure 1: Example of olog.

3 Semi-automated Olog Construction using Text Generation AI

To explore the potential of semi-automated database schema (olog) construction, we applied ChatGPT to assist in generating an olog from the credit recognition regulations of Institute of Science Tokyo. The goal was to evaluate whether a text generation AI support key steps such as concept extraction, relationship definition, and compositional verification, while reducing the need for extensive manual effort. The final result is presented in Figure 2.
Before constructing the olog, we carried out several preparatory steps:

I. We instructed ChatGPT to respond as an expert in category theory and database design.
II. We provided papers [1,2,3,4,5] on categorical databases and ologs, and requested to build its background knowledge. We then requested detailed explanations of these references, and where responses were unclear or incomplete, we followed up with further questions to refine its understanding.
III. We requested ChatGPT to explain the process of olog construction, placing particular emphasis on the requirement that all morphisms must be functional, and that the existence and semantic validity of composite morphisms must be carefully verified.

These preparatory steps provide the basic knowledge needed for constructing a categorical database. When building a database from natural-language institutional documents, it is important that all participants share a clear understanding of category theory and the semantics of ologs. By following these procedures, ChatGPT was able to act not just as a content generator, but also as a collaborative partner that helped interpret domain-specific language and organize it into a database schema with mathematical consistency.

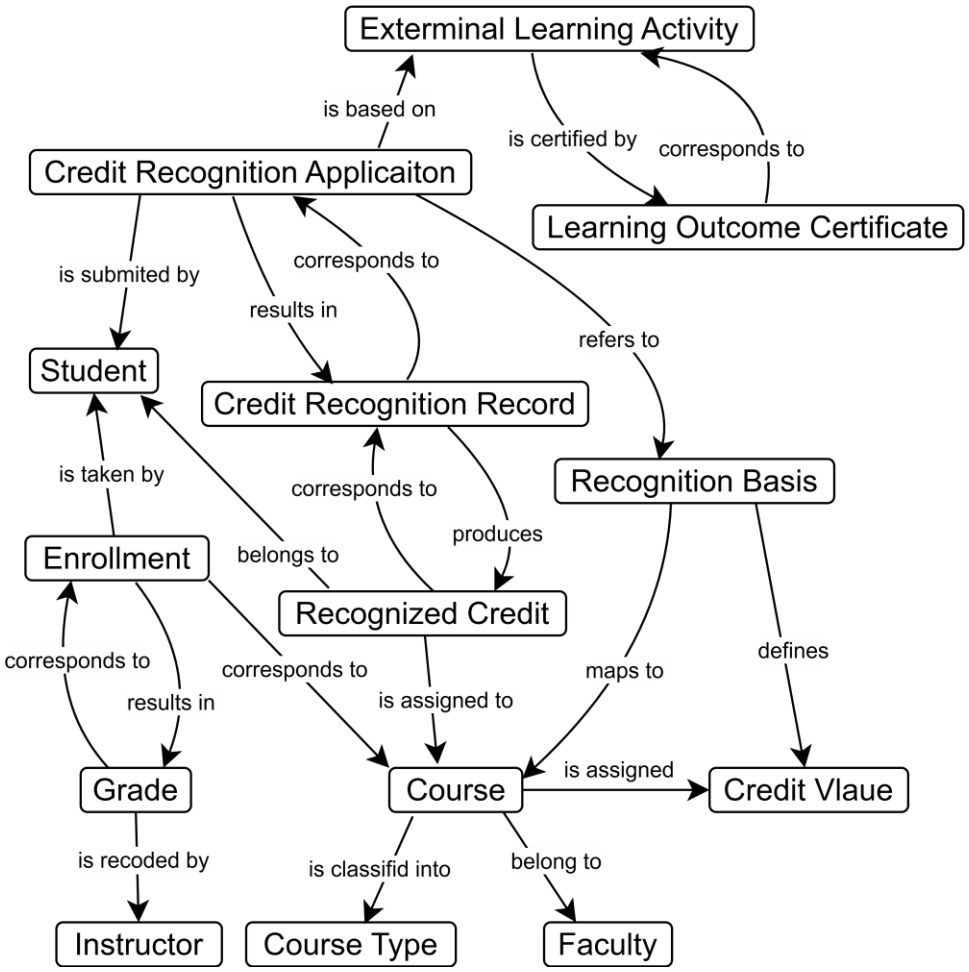

Figure 2: Olog for University Course Management.

1. Extracting Database Concepts from the Academic Regulations

We began by directing ChatGPT to analyze the Academic Regulations of Institute of Science Tokyo, with particular attention to the provisions concerning credit recognition. After providing the official document, we guided ChatGPT to extract and list key entities that should be represented as database tables—namely, concepts or objects—to support operations related to academic credit recognition.

As a complementary step, we also asked ChatGPT to refer to the credit recognition guidelines issued by Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). These national-level guidelines, which often serve as foundational references for university regulations, were expected to help clarify and complete the identification of relevant concepts. ChatGPT identified 14 concepts to be modeled as database tables within the categorical schema for credit recognition:

[Student], [Instructor], [Faculty], [Course], [Course Type], [Credit Value], [Enrollment], [Grade], [Credit Recognition Application], [Recognized Credit], [Recognition Basis], [Credit Recognition Record], [Learning Outcome Certificate], and [External Learning Activity].

These concepts formed the foundational object set for defining morphisms and examining composability in accordance with categorical database design principles.

2. Formalizing Conceptual Relationships and Verifying Functionality

After identifying the 14 concepts from the academic regulations, we moved on to formalizing the relationships between concepts as natural language statements. In an olog, each morphism must represent a functional relationship, meaning it must satisfy both totality (the relationship applies to all instances) and uniqueness (each instance maps to exactly one target). Functionality corresponds to one-to-one or many-to-one relationships in database modeling.
We instructed ChatGPT to propose candidate relationships that appeared to be functional, expressed as simple sentences. For one-to-one relationships, we requested bidirectional formulations to capture the mutual correspondence. Through this process, 21 relationships connecting the 14 identified concepts were established. Examples include:

- [Enrollment] results in [Grade]
- [Course] belongs to [Faculty]
- [Grade] is recorded by [Instructor]

For each relationship, we asked ChatGPT to justify whether it satisfied uniqueness, drawing on institutional definitions, document structures, terminology, and domain-specific reasoning. In many cases, uniqueness was directly supported by institutional semantics. For example:

- "[Recognized Credit] is awarded to [Student.]"
  → The regulations presume that recognized credits cannot exist independently of a student,
  → satisfying uniqueness.

In other cases, explicit confirmation from the regulations was lacking. However, ChatGPT proposed reasonable assumptions to maintain functionality, allowing schema designers to validate them efficiently with minimal effort.
Next, we assessed whether each relationship satisfied totality—that is, whether the mapping covered all instances in the domain. ChatGPT analyzed each relationship based on the regulatory structure and the institutional logic. The results were:
18 out of 21 relationships were found to satisfy totality, supported by clear institutional rules. Examples include:

- "[Enrollment] is taken by [Student.]"
  → An enrollment record must always belong to a student.
  → satisfying totality.
- "[Course] belongs to [Faculty.]"
  → Every course must be offered by a faculty or department."
  → satisfying totality.

Two relationships appeared to satisfy totality conditionally and could likely be confirmed through further institutional review. One relationship could not be verified with the available information and would require explicit documentation or policy inspection.
Overall, although fully automating the verification of relationships remains challenging,

ChatGPT supported the process. It not only proposed semantically consistent relationships but also provided reasoned assessments of functionality and totality, greatly reducing the manual effort required for schema validation.

3. Composition of Successive Morphisms

After establishing a complete set of functional relationships between concepts (as shown in Figure 2), we had a preliminary olog schema. However, this structure alone was not sufficient to constitute a fully valid categorical database. To meet the categorical definition, it was necessary to verify that composite morphisms exist and are semantically coherent.

In an olog, composite morphisms correspond to sequences of functional relationships connecting concepts through intermediate steps. Verifying a composite means checking whether the sequence can be replaced by a single morphism that preserves both the meaning and interpretive content of the original chain. This requirement is not only a formal constraint from category theory (composability) but also a semantic necessity in olog design, ensuring that multi-step relationships can be expressed as coherent and declarable facts.

To carry out this verification, we asked ChatGPT to enumerate all meaningful composite paths among the 14 core concepts. It identified 18 paths of length 2 (three concepts), 11 paths of length 3, and 3 paths of length 4, for a total of 32 composite paths. For each, we requested a natural language sentence that would express the composite morphism as a declarative and semantically faithful statement.

Each generated sentence was evaluated based on four criteria:

- Syntactic alignment: The starting and ending concepts match those of the original path.
- Semantic inclusion: The meaning of intermediate concepts is implicitly reflected within the sentence.
- Linguistic compactness: The sentence is concise and not a verbose concatenation of individual steps.
- Semantic equivalence: The sentence has the same interpretation as the original chain and can be declared as a fact in the olog.

Examples include:

Consider the path, [Grade] → [Enrollment] → [Student]. This sequence can be logically expanded as

"[Grade] corresponds to [Enrollment,] and [Enrollment] is taken by [Student.]"

ChatGPT proposed the composite sentence:

"[Grade] is given to [Student.]"

This formulation preserves the syntactic start and end points while naturally embedding the meaning of [Enrollment] — implying that grades are awarded only in the context of a student's enrollment. The sentence is linguistically compact and semantically equivalent to the original path, making it a well-formed composite morphism in the olog.

Across all paths, we confirmed that valid and semantically equivalent composites could be expressed as single natural language sentences. This indicates that the constructed olog schema satisfies the categorical requirement of composability and supports human-readable expressions of data semantics.

Verifying the semantic validity of composite morphisms is inherently complex and requires careful interpretation. Despite this, our guided process enabled efficient verification, as described in the following section. Whether a composite preserves the meaning of its steps often depends on domain knowledge and contextual understanding. While ChatGPT cannot fully guarantee semantic equivalence, it was highly effective in proposing candidate expressions and greatly accelerated the verification process.

4 Conclusion

Constructing structured databases directly from natural-language institutional documents is a challenging task. It usually requires both a deep understanding of the subject matter and a lot of manual effort. To make this easier, we explored how a text generation AI (ChatGPT) could help in constructing categorical databases, especially olog schemas. We simplified database design by helping to extract key concepts, define consistent relationships, and verify the overall structure based on category-theoretic ideas. Through a guided process, we showed that semi-automated schema building is not only possible but also effective, even for users without specialized knowledge.

We led the language model through a step-by-step workflow. It started with extracting key concepts from academic regulations, then moved to describing relationships in natural language, and finally checked whether composite relationships could be properly formed. As a result, we created an olog schema that satisfied the requirements of categorical database design, including functionality and composability. ChatGPT was able to generate clear and meaningful relationships, help reason about their functional properties like uniqueness and totality, and assist in validating multi-step compositions by expressing them in compact and understandable natural language.

One important outcome of this approach is that it allows people without deep expertise in category theory or database design to work on schema building. Through conversations with the AI, even those without formal training could review and validate relationships and better understand the structure of institutional rules. This suggests a promising way to make categorical database modeling more accessible, especially in educational and administrative fields where data integrity matters but modeling expertise is often limited.

The text generation AI identified 14 academic concepts and 21 functional relationships from the university's credit recognition regulations. We also examined 32 composite morphisms and confirmed their semantic validity using simple and concise natural language summaries. The AI's ability to represent complex relationships in clear and natural language made the overall database structure much easier to understand and communicate.

This work demonstrates that integrating ologs with text generation AI offers a practical and powerful method for building structured databases in a semi-automated way. It reduces the manual burden on designers while improving the clarity and maintainability of data systems. Future work may expand this approach to cover more types of documents, introduce deeper reasoning abilities, and allow dynamic updates to schemas. Overall, our study demonstrates that AI can meaningfully help formalize institutional logic into structured, analyzable data models.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP22H00077.

References

[1] D.I. Spivak, "Functorial data migration," Information and Computation 217, 2012, pp. 31-51.

[2] D.I. Spivak, R. Wisnesky, "Relational Foundations for Functorial Data Migration," arXiv:1212.5303v7 [cs.DB] 24 Jul 2015, 2015.

[3] D.I. Spivak, R.E. Kent, "Ologs: A categorical framework for knowledge representation," arXiv:1102.1889v2 [cs.Lo] 7 aug 2011, 2011.

[4] T. Inakura, S. Imai, K. Takamatsu, S. Matsumoto, M. Mori, "Application of Category Theory to Database Construction in Institutional Research," Proceedings of the 12th Meeting on Japanese Institutional Research, 2023, pp. 170-175, in Japanese.

[5] T. Inakura, S. Imai, K. Takamatsu, S. Matsumoto, M. Mori, "A database that ensures data integrity based on category theory," Proceedings of the 40th Annual Meeting of Japanese Society of Educational Information, 2024, pp. 66-69, in Japanese.