Journal of Artificial Intelligence Practice (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2026.090203
ISSN 2371-8412 Vol. 9 Num. 2

Construction and Question-Answering Application of Research Management Knowledge Graph for Multi-Source Heterogeneous Data — A Case Study of Research Application Qualification

Yang Lei

Yunnan Open University, Kunming, China
271092522@qq.com

Keywords: Knowledge graph; Multi-source heterogeneous data; Research management; Intelligent question-answering; Retrieval-augmented generation

Abstract: Research management in universities involves massive multi-source heterogeneous data scattered across diverse systems and formats. Conventional relational database management modes can hardly satisfy the requirements of cross-domain associative query and intelligent question-answering. To tackle this practical challenge, this paper takes university research management as the application scenario and proposes a scheme for constructing a research-management knowledge graph and its question-answering application oriented to multi-source heterogeneous data. This study systematically sorts out the heterogeneous features of data sources in research management, designs a knowledge-graph construction workflow covering data access, ontology modeling, knowledge extraction, knowledge fusion and knowledge storage. Furthermore, a knowledge-graph question-answering framework integrating semantic parsing and graph query is built. Combined with retrieval-augmented generation technology, interpretable intelligent question-answering services are realized. Taking the qualification review for research applications at Yunnan Open University as the validation scenario, experimental results demonstrate that the proposed method can effectively integrate multi-source data including title evaluation, project application and achievement recognition. It improves associative query efficiency and question-answering accuracy, and provides a feasible technical path for the intelligent transformation of university research management.

1. Introduction

With the growing volume of research activities and continuous increase of research investment in universities, the scale and complexity of data faced by research management rise exponentially. Business procedures such as title evaluation, project application, achievement recognition, fund management and team building generate large quantities of scattered information. Such information is stored in multiple independent platforms including research-management information systems, personnel systems, financial systems and educational administration systems, resulting in the "information island" problem. Different systems adopt disparate data formats, coding standards and

semantic conventions without unified association mechanisms. Consequently, research administrators are hindered from conducting cross-system integrated queries and in-depth analysis, while individual teachers cannot quickly obtain a complete view of their own research status.

Knowledge graphs organize entities and their semantic relations in graph structures, offering an effective approach for knowledge-oriented organization and associative reasoning of massive heterogeneous data[1]. Academia has formed a relatively systematic understanding of the technical system for knowledge-graph construction, covering key links such as knowledge representation, knowledge extraction, knowledge fusion and knowledge reasoning[1]. In research domains, graph-based representation of core entities (researchers, institutions, achievements, projects, etc.) and their relations lays a data foundation for research situation awareness and knowledge services[2]. Meanwhile, knowledge graphs have been widely adopted in various scenarios of industry-oriented knowledge services. Their advantages in complex associative queries as semantic networks have been fully verified[3].

Nevertheless, existing studies mostly focus on knowledge-graph construction for large-scale scientific literature data or single industrial domains. Research on knowledge-graph construction and question-answering applications targeting multi-source heterogeneous data for university research management — a specific scenario characterized by strong constraints and high timeliness — remains insufficient. Research-management data features diverse sources and heterogeneous structures, as well as complicated business rules and frequent updates. It imposes higher requirements on the accuracy of knowledge extraction, consistency of knowledge fusion and interpretability of question-answering systems. Given the above context, this paper conducts research on knowledge-graph construction for multi-source heterogeneous data and intelligent question-answering in university research-management scenarios, attempting to address three core questions: first, how to implement ontology modeling and knowledge extraction for multi-source heterogeneous research-management data; second, how to realize entity alignment and knowledge fusion for cross-source data; third, how to build a knowledge-graph question-answering application with both accuracy and interpretability.

2. Analysis on Characteristics of Multi-Source Heterogeneous Data in Research Management

2.1. Diversity and Heterogeneity of Data Sources

University research-management data originate from extensive sources: personnel information, project information and achievement information in research-management information systems; title and post information in personnel systems; fund information in financial systems; teaching task information in educational administration systems, and so forth. These data sources differ significantly in storage media, data structures, update cycles and semantic granularity, forming a typical heterogeneous data environment[4]. In terms of structurization level, research-management data include highly structured data in relational databases, semi-structured documents such as project application forms and project closure reports, as well as unstructured texts including policy documents and official announcements, presenting prominent multi-source heterogeneous features[4]. Table 1 summarizes peripheral system data sources and their heterogeneous features involved in decision-making and management within research management.

Table 1 Data sources and heterogeneous features of research management

Data Source	Data Content	Structurization Level
Personnel System	Professional title, educational background, age, admission time, post	Structured
Research	Ongoing projects, closed projects,	Semi-structured

Management System	achievements, funds	
Financial System	Project fund usage, indirect costs	Structured
Policy Documents	Project guidelines, application measures, title evaluation documents	Unstructured

From the perspective of data structure, different systems often employ distinct table structures, field naming conventions and coding specifications. The same entity (e.g., the same teacher or research project) may be represented inconsistently across different systems, which renders entity disambiguation and alignment an inevitable requirement for knowledge fusion. Semantically, research management involves numerous domain terms related to professional titles, project categories, achievement types and fund subjects. The connotation and denotation of these terms vary under different contexts, calling for ontology-based semantic standardization and unification.

2.2. Complexity and Dynamics of Business Rules

Business rules for research management are highly complex and dynamic. Taking project application qualification as an example, different types of projects impose differentiated constraints on applicants' professional titles, educational backgrounds, ages, preliminary achievements and the number of ongoing projects. Such constraints change dynamically alongside policy adjustments. Title evaluation requires comprehensive judgment on multi-dimensional indicators covering teaching, research and social services. Achievement recognition covers multiple types and their hierarchical classifications, including papers, monographs, patents, awards and decision-making consultation reports. These rules are massive in quantity, nested with one another and highly coupled with specific data, which cannot be flexibly expressed via conventional relational queries.

The above characteristics indicate that the construction of research-management knowledge graphs cannot simply adopt generic knowledge-graph construction workflows. Domain features must be fully considered in data modeling, knowledge extraction and knowledge fusion. For question-answering applications, precise matching of rule constraints and semantic understanding for natural-language interaction should both be taken into account.

3. Construction Methodology for Research-Management Knowledge Graph

3.1. Ontology Modeling and Schema-Layer Design

Ontology modeling serves as the foundation for knowledge-graph construction. Its core task is to define core concepts, attributes and mutual relations within the research-management domain so as to form the schema layer of the knowledge graph. Combined with practical research-management business, this study defines core entity types including researchers, research projects, research achievements, research institutions, academic activities, fund management and evaluation indicators. Semantic relations such as undertake, produce, hold position, preside over, participate in and correspond to are established. For instance, a researcher is linked to research projects via the undertake relation, linked to research achievements via the produce relation, and linked to research institutions via the hold position relation. Projects and achievements are associated with each other through the rely on relation.

Schema-layer design should balance universality and scalability. Universality requires core concepts and relations to cover major business domains of research management; scalability demands the ontology structure to accommodate new concepts and relations derived from emerging businesses such as title evaluation and project application. Therefore, this research adopts a hierarchical modeling strategy of "core ontology + domain extension". On the basis of a stable core framework, extension modules are utilized to adapt to continuously changing business requirements.

Taking research application qualification as an example, ontology acts as the skeleton of the knowledge graph. The schema layer is built in a top-down manner, while the data layer is populated bottom-up[6].

3.1.1. Design of Six Core Entities

Researcher: professional title, educational background, age, field of expertise, admission time, number of ongoing projects

Project: project ID, name, category, fund, start-end time, status

ProjectCategory: e.g., National Social Science Fund, National Natural Science Foundation, Yunnan Philosophy and Social Science Project, Education Department Project

Achievement: papers, monographs, patents, awards, research reports

Department: faculties, Research Office, Finance Office

QualificationRule: formal description of application conditions

3.1.2. Semantic Analysis of Entities and Framework Design of Knowledge Graph

Key semantics of the above-mentioned entities in the knowledge graph are sorted out in Table 2.

Table 2 Semantic relations among core entities for research application qualification

Relation	Head Node	Tail Node	Semantics	Key Attributes
apply for	Researcher	Project	A teacher applies for a certain project	Application time, status, result
belong to category	Project	ProjectCategory	Category that a project belongs to	—
have title	Researcher	Title	Current professional title	Rank, acquisition time
have education	Researcher	Education	Highest educational attainment	Degree, awarding institution, time
produce achievement	Researcher	Achievement	Achievements obtained by a teacher	Publication time, rank, authorship
preside over ongoing	Researcher	Project	Ongoing research projects	Start-end time, fund
affiliate with	Researcher	Department	Personnel affiliation	Post type
apply to	ProjectCategory	QualificationRule	Applicable conditions for a certain project category	Condition type, constraint value
restrict	QualificationRule	Researcher	Constraints for non-compliance	Description of missing items

The knowledge-graph relation design for application qualification in research management is illustrated in Figure 1.

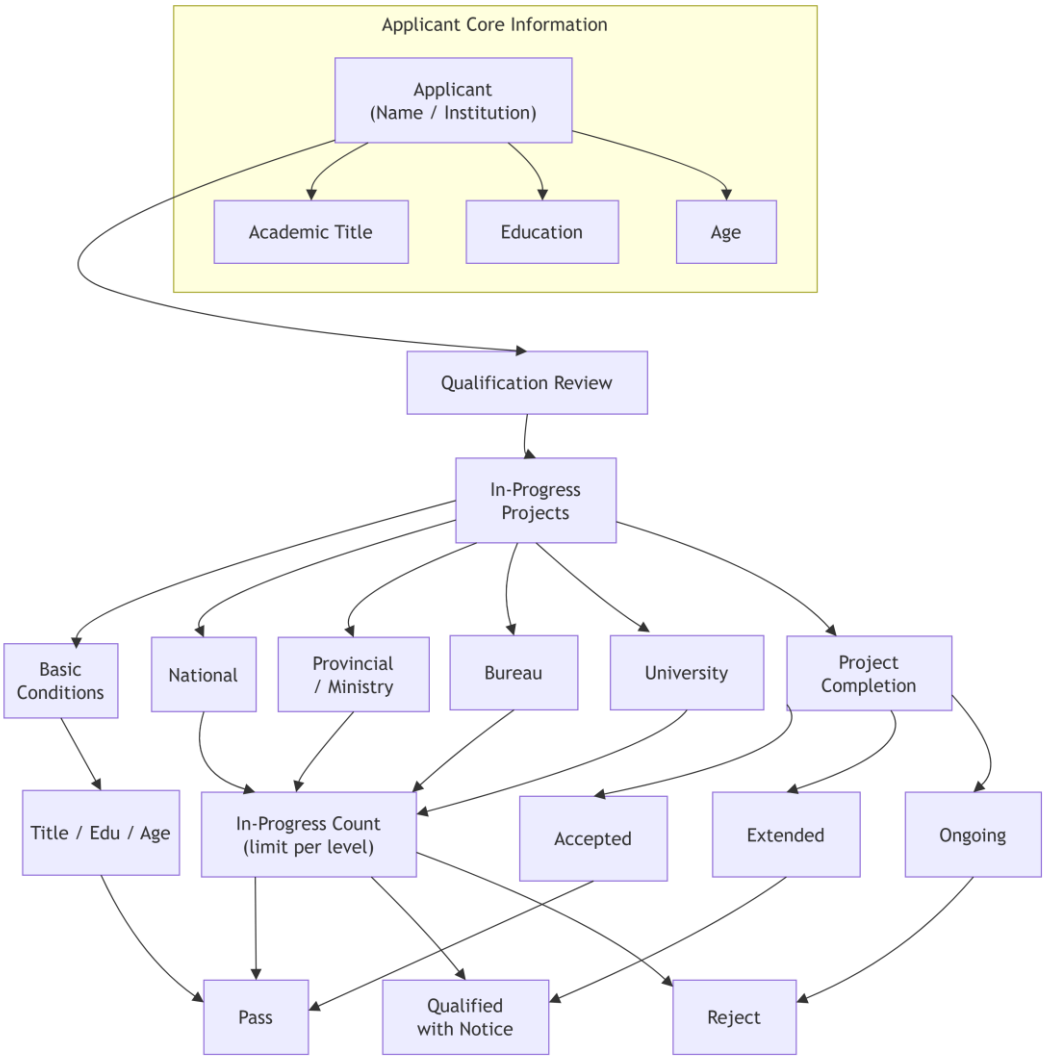


Figure 1 Knowledge-graph structure for research application qualification review, reference to the original diagram

3.1.3. Graph Modeling for Qualification Rules

Qualification conditions can be converted into constraint edges with attributes. For example, the rule “Applicants for general projects of the National Social Science Fund shall hold an associate-senior professional title or a doctoral degree” can be formally expressed as:

(:ProjectCategory {name:"General Project of National Social Science Fund"})

-[:apply_to {dimension:"professional title", operator:">=", threshold:"associate-senior"}]->

(:QualificationRule {ruleID:"R001"})

Such formal representation enables direct reading by rule engines (e.g., Drools), realizing the division of labor: “knowledge graph stores relations, and rule engine performs judgment”[7].

3.2. Multi-Source Data Access and Knowledge Extraction

Access and knowledge extraction of multi-source heterogeneous data constitute major difficulties in knowledge-graph construction. For structured data, this study adopts data mapping to transform tables in relational databases into entities and relations in the graph, and realizes automatic data

conversion through field-level mapping rules. For semi-structured data, template matching and rule parsing are combined to extract structured elements from documents. For unstructured texts, entity-relation extraction techniques are adopted to identify entities such as researchers, projects and achievements as well as their semantic relations.

Entity-relation extraction is a critical technique for knowledge-graph construction, whose accuracy directly determines graph quality[6]. Traditional approaches rely on manually constructed feature templates and lexicons, which entail heavy workload and limited generalization capability. Deep-learning-based extraction methods can automatically learn semantic features from texts and have achieved remarkable performance in general domains. Nevertheless, they still face challenges such as lack of annotated data when applied to research-management texts characterized by specialized terminology and standardized sentence patterns[6]. Accordingly, this study introduces zero-shot and few-shot extraction capabilities of large language models on the basis of rule-based extraction. Prompt engineering for research-management policy texts is designed to implement automatic entity-relation extraction, supplemented by manual review mechanisms to guarantee extraction quality[13]. Practical research demonstrates that joint entity-relation extraction methods for domain-specific knowledge graphs exhibit prominent advantages in handling overlapping relations and scenarios with insufficient annotated samples, providing methodological references for extraction tasks in research-management domains[8].

3.3. Cross-Source Knowledge Fusion and Entity Alignment

Knowledge fusion addresses the problem of multiple representations for identical entities from multi-source data, realizing entity alignment and conflict resolution for cross-source knowledge. In research-management scenarios, the same teacher may be recorded with different name formats or staff IDs across different systems; the same research project may be described differently in various databases. Therefore, attribute matching, semantic similarity calculation and graph-structure constraints are adopted for entity disambiguation[5].

This research implements knowledge fusion with comprehensive strategies as follows:

Entity alignment based on attribute similarity: similarities are calculated by comparing key entity attributes (staff ID, full name, ID number, project ID, etc.), and thresholds are set to judge whether entities refer to the same object.

Graph-structure constraints combining context and adjacent relations: associated information under special cases such as “same name for different persons” and “different names for the same person” assists disambiguation.

Conflict resolution mechanism: for attribute-value conflicts from diverse data sources, judgments are made according to data authority and timeliness to ensure consistency and reliability of fused knowledge.

Essentially, knowledge fusion for multi-source heterogeneous data unifies associated knowledge distributed in disparate data sources into one integrated graph structure, so as to provide consistent data views for cross-system associative queries[5,7].

3.4. Knowledge Storage and Update Mechanism

Upon knowledge-graph construction completion, an appropriate graph database should be selected for knowledge storage. This study chooses Neo4j as the storage and query engine for the knowledge graph. Relying on its property-graph model and Cypher query language, it realizes flexible organization and efficient retrieval of research-management entities and relations. Centered on nodes and edges, the data organization mode of graph databases naturally fits graph structures of knowledge graphs. It delivers superior performance compared with traditional relational databases in multi-hop

associative queries, path analysis and subgraph extraction.

Considering frequent updates of research-management data, an incremental knowledge update mechanism is established. By configuring data-source change detection and regular synchronization strategies, additions, modifications and deletions of source data are transformed into corresponding operations within the knowledge graph to implement dynamic maintenance. Meanwhile, a version-management mechanism is built to record change trajectories across graph versions and support data traceability and historical analysis.

4. Knowledge-Graph-Based Question-Answering Application for Research Management

4.1. Overall Framework of Knowledge-Graph Question-Answering

After knowledge-graph construction, this study designs and implements a knowledge-graph question-answering system oriented to research management. Its overall framework consists of four modules: query understanding, graph retrieval, answer generation and evidence traceability. The query-understanding module parses natural-language user questions into executable query intents. The graph-retrieval module performs retrieval and reasoning on the knowledge graph according to parsing outputs to obtain relevant entities and relations. The answer-generation module organizes graph-retrieval results into natural-language answers. The evidence-traceability module presents knowledge paths supporting answers for users to guarantee answer interpretability.

In terms of technical routes, the system adopts a dual-channel strategy integrating “semantic parsing + graph query” and retrieval-augmented generation. For rule-based questions with definite intents and clear structures (e.g., whether a teacher meets application qualifications for a certain project category), semantic parsing converts natural-language questions into Cypher statements for precise matching on the graph, generating answers with rigorous logical foundations. For consultative open-ended questions requiring synthesis of multiple knowledge fragments, retrieval-augmented generation is adopted: relevant evidence fragments are retrieved from the knowledge graph and knowledge base, then organized and generated by large language models, balancing response fluency and factual accuracy[9,12].

4.2. Intent Recognition and Query Parsing

Intent recognition is the primary step for question-answering systems. Targeting research-management question-answering scenarios, this study establishes an intent system covering qualification inquiry, policy consultation, progress inquiry, achievement retrieval and relation discovery. A hybrid approach combining rule-based and semantic-understanding-based methods is utilized for intent recognition. For questions matching predefined templates (e.g., “What are the application conditions for Project XX?”), direct mapping is performed to corresponding intents. For free-text questions with flexible expressions, large-language-model capabilities are leveraged for intent classification and slot filling.

At the query-parsing stage, entities and relations identified are mapped to the schema layer of the knowledge graph to generate structured queries. For instance, given the query “Which ongoing projects is Teacher Zhang undertaking?”, the system recognizes the researcher entity “Teacher Zhang” as well as relations and attribute constraints corresponding to “undertake” and “ongoing”, then generates Cypher queries to acquire qualified project nodes. For complex questions involving multi-hop associations, path planning enables hop-by-hop graph retrieval to realize associative reasoning across entities[10].

4.3. Interpretable Question-Answering and Evidence Traceability

Interpretability serves as the core requirement distinguishing research-management question-answering from generic question-answering. Scenarios including research-qualification judgment and achievement recognition require systems to provide not only conclusions but also supporting rules and data chains. This requirement is fulfilled by semantic paths inherent to knowledge graphs: while generating answers, the system records traversed entities, relations and attribute constraints during queries, and demonstrates reasoning processes for users in the form of graph paths[10]. When judging whether a teacher satisfies the application qualification for a certain project category, for example, the complete association path "researcher — professional title — project" together with judgment bases of each link will be displayed, rendering conclusions verifiable and traceable.

This mechanism effectively mitigates the "black-box" defect of pure large-language-model generation. Existing research indicates that combining knowledge graphs with large language models — adopting knowledge graphs to provide verifiable factual skeletons and assigning language organization tasks to large models — can significantly reduce hallucination risks while improving the interpretability and reliability of question-answering[9,11]. Following this technical path, this study realizes complementary advantages of structured reasoning of knowledge graphs and semantic generation of large language models[11,12].

5. Conclusion

Constructing knowledge graphs and developing question-answering applications for multi-source heterogeneous data represent a vital technical route for the digital transformation of university research management. This paper systematically analyzes the diversity and heterogeneity of research-management data, designs a knowledge-graph construction workflow covering ontology modeling, knowledge extraction, knowledge fusion and knowledge storage, and builds a knowledge-graph question-answering framework integrating semantic parsing and retrieval-augmented generation. Validation experiments are conducted with Yunnan Open University as the application scenario. Research results show that knowledge graphs can effectively integrate research-management data scattered in disparate systems. Knowledge-graph-powered question-answering delivers accurate and interpretable intelligent services for teachers and administrators, offering feasible solutions and reference examples for intelligent upgrading of university research-management services.

Acknowledgement

Key Project of Yunnan Open University University-level Fund: Research and Application of RAG-based Intelligent Question-Answering System — Taking Application Qualification Verification in Scientific Research Management as an Example（25YNOU05）.

References

[1] Liu Q, Li Y, Duan H, Liu Y, Qin Z G. A Survey on Knowledge Graph Construction Techniques[J]. Journal of Computer Research and Development, 2016.

[2] Zhou Y C, Wang W J, Qiao Z Y, Xiao M, Du Y. Review on Construction Methods and Applications of Knowledge Graph for Scientific Big Data[J]. Science China: Information Sciences, 2020.

[3] Huang H Q, Yu J, Liao X, Xi Y J. Review of Research on Knowledge Graph[J]. Computer Systems & Applications, 2019.

[4] Jiang B C, Wan G, Xu J, Li F, Wen H Q. Construction of Large-Scale Geographic Knowledge Graph with Multi-source

Heterogeneous Data[J]. Acta Geodaetica et Cartographica Sinica, 2018.
[5] Wang B R, Zhou X Z, Wei W, Wang R, Liu Y M, et al. Construction and Application of Neurosurgery Knowledge Graph Based on Multi-source Heterogeneous Data[J]. Journal of Beijing Institute of Technology, 2024.
[6] E H H, Zhang W J, Xiao S Q, Cheng R, Hu Y X, Zhou X S, et al. Review of Entity-Relation Extraction Based on Deep Learning[J]. Journal of Software, 2019.
[7] Zhang X S, Liu L, Wang H L, Su G B, et al. Research on Entity-Relation Extraction Methods in Knowledge Graph[J]. Journal of Computer Science and Exploration, 2024, 18(3).
[8] Fu R, Li J Y, Wang J H, Yue K, Hu K. Joint Entity-Relation Extraction for Domain-Specific Knowledge Graphs[J]. Journal of East China Normal University (Natural Science), 2021.
[9] Ma Z G, Ni R Y, Yu K H. Recent Advances, Key Technologies and Challenges of Knowledge Graph[J]. Journal of Engineering Science, 2020. DOI:10.13374/j.issn2095-9389.2020.02.28.001.
[10] Lin W Q, Wang M, Wang W, Wang C N, Jin S C. Research Status and Military Applications of Knowledge Graph[J]. Journal of Chinese Information Processing, 2021.
[11] Huang B, Wu S A, Wang W G, Yang Y, Liu J, et al. Graph-Model Complementarity: A Review on the Fusion of Knowledge Graphs and Large-scale Models[J]. Wuhan University Journal (Natural Science Edition), 2024.
[12] Zhang K L, Wang Y, Fu W H, Zhu Y Q, Zhang Y L, et al. Review on Construction and Application of Knowledge Graph Driven by Large Language Models[J]. Journal of Zhengzhou University (Natural Science Edition), 2026.
[13] Zhang C K, Li X L, Zheng S, Cai J J, et al. Research on Construction and Application of Knowledge Graph Based on Large Language Models[J]. Journal of Computer Science and Exploration, 2024, 18(10).