KnowOntoAsk: An Ontology-Based Framework for Human-AI Collaborative Knowledge Exploration in Organizational Decision-Making Processes

Antony Seabra de Medeiros, Daniel Schwabe, Sergio Lifschitz

[1]Departamento de Informatica - PUC-Rio

amedeiros, sergio@inf.puc-rio.br, dschwabe@gmail.com

Abstract. *Knowledge exploration to support decision-making is inherently challenging, as individuals often begin this process with imprecise questions. Moreover, it depends on the ability to explore information distributed across heterogeneous sources and, compounding the challenge, past organizational decisions are rarely reusable, resulting in the loss of valuable institutional knowledge that could inform future choices. This work proposes KnowOntoAsk, a knowledge-driven framework that leverages formal ontologies to model domain knowledge across diverse data sources and guide interactive, goal-oriented dialogues. The framework includes an onboarding phase to help individuals articulate their initial objectives, and the generation of context-aware clarifying questions to refine their queries and uncover missing information, enabling more effective discovery and better-informed decision-making in organizations.*

Resumo. *A exploração de conhecimento para apoiar a tomada de decisão é, por natureza, desafiadora, uma vez que indivíduos frequentemente iniciam esse processo com perguntas imprecisas. Além disso, ela depende da capacidade de explorar informações distribuídas em fontes heterogêneas e, para agravar o desafio, decisões organizacionais passadas raramente são reutilizáveis, resultando na perda de conhecimento institucional valioso que poderia orientar escolhas futuras. Este trabalho propõe o KnowOntoAsk, um framework orientado por conhecimento que utiliza ontologias formais para modelar o domínio a partir de fontes de dados diversas e conduzir diálogos interativos orientados a objetivos. O framework inclui uma fase de onboarding, que auxilia os indivíduos na formulação de seus objetivos iniciais, e a geração de perguntas de esclarecimento sensíveis ao contexto, com o objetivo de refinar suas consultas e revelar informações ausentes — promovendo, assim, uma descoberta mais eficaz e decisões mais bem fundamentadas nas organizações.*

1. Academic Information

Level	PhD
Advisor	Sergio Lifschitz
Coadvisor	Daniel Schwabe
Admission	2022-2
Foreign Language Proficiency	2022-2
All Credits Completed	2024-1
Qualifying Exam	2024-2
Thesis Proposal Defense Expected	2025-1
Final Defense Expected	2025-2

2. Introduction

The notion that *a problem well put is half solved*, as articulated by [Dewey 1910], underscores the importance of well-formulated questions for effective problem-solving and informed decision-making. This perspective emphasizes that the ability to craft precise and insightful questions is just as vital as the ability to provide accurate answers. In contrast, people often struggle to articulate their information needs clearly, especially in complex or unfamiliar domains. This is understandable, as long as individuals typically progress through their academic careers by providing answers, from primary school all the way through high school. Tests are primarily designed to gauge what students know, rather than to encourage them to formulate their own questions. This emphasis on answering rather than questioning extends to modern AI tools like ChatGPT and Gemini, which are largely focused on generating responses to user prompts.

While the ability to ask better questions is fundamental, it is equally important to ensure that the answers are grounded in comprehensive and relevant knowledge. In organizational contexts, information is scattered across a wide range of heterogeneous sources, including databases, unstructured documents, spreadsheets, email threads, among others. These sources often use different formats, terminologies, and levels of structure, creating semantic and technical barriers to integration. As a result, even when individuals manage to formulate meaningful questions, retrieving consistent and complete answers remains a challenge. Effective knowledge exploration in organizations therefore demands not only the ability to express intent, but also mechanisms to semantically bridge and unify these diverse sources into a coherent and accessible knowledge space.

Beyond data integration, another critical aspect of organizational knowledge is its temporal dimension: how decisions and insights from the past are captured, stored, and reused. In practice, many organizations lack structured processes to document their rationale for making decisions or the contextual factors that influenced past outcomes.However, leveraging this memory requires more than archiving reports; it demands a framework that connects current information needs with past decisions in a meaningful way. By embedding historical insights into the knowledge exploration process, organizations can shift from reactive, case-by-case decision-making to a more informed, cumulative, and strategic approach.

3. Research Problem

Organizations increasingly rely on data to support decision-making, yet their information is often dispersed across heterogeneous sources—ranging from structured databases and spreadsheets to unstructured documents, reports, and communication platforms. This fragmentation makes it difficult to form a unified view of organizational knowledge and limits the ability of individuals to explore and make sense of available information. Moreover, individuals frequently struggle to articulate precise and context-aware questions, especially in complex or unfamiliar domains, and often lack clarity on how to begin the exploration process. In many cases, even the objectives driving the inquiry remain implicit or poorly defined. Compounding this challenge is the fact that decisions, once made, are rarely documented in a structured, reusable format—leading to the loss of valuable organizational memory that could inform future actions. This gives rise to the following central research problem.

How can organizations integrate their distributed information assets and accumulated decision history to support individuals in initiating and refining knowledge exploration, ultimately enabling more informed and context-aware decision-making?

Addressing this challenge requires not only semantic integration and retrieval across data silos, but also interaction models that guide individuals from onboarding, helping them express their intent, all the way through progressive clarification and contextualization. Based on these gaps, the research focuses on the following key problem areas:

- *Designing a conceptual architecture* for knowledge exploration in decision-making contexts, integrating knowledge bases derived from heterogeneous data sources and supports onboarding, clarification, and organizational memory.
- *Enabling knowledge exploration through dialogue*, where users can iteratively refine their queries via context-aware clarifying questions grounded in formal knowledge representations.
- *Preserving organizational memory* by capturing and reusing past decisions, thereby supporting institutional learning and avoiding knowledge loss across time and teams.

4. Related Work

The study of Clarification Questions (CQs) has gained significant attention across various domains, particularly in Natural Language Processing (NLP), Information Retrieval (IR), and Human-Computer Interaction (HCI). CQs are necessary for resolving ambiguities, filling information gaps, and improving user satisfaction in interactive systems. Early work in dialogue systems recognized the importance of clarification in human-human communication to resolve misunderstandings and maintain dialogue flow [Ginzburg 2001, Purver et al. 2003]. These foundational insights have driven research into developing computational models for generating effective CQs.

In the realm of IR, [Zamani et al. 2020] proposed models for generating clarifying questions in web search, demonstrating their positive impact on clickthrough rates. More recently, [Eberhart and McMillan 2022] explored CQ generation for source code search, highlighting their utility in query refinement for technical domains. The effectiveness of CQs in IR has also been investigated from the perspective of user satisfaction, with studies emphasizing the importance of specific and well-formulated questions to avoid user frustration [Aliannejadi et al. 2024a, Aliannejadi et al. 2024b]. These works underscore that not all CQs are equally beneficial, and understanding the features that contribute to their usefulness is critical.

Within NLP, research on CQs often focuses on their generation and ranking. [Rao and Daumé III 2018] introduced retrieval and generation models for CQs, with a focus on controlling the specificity of generated questions to elicit useful information. [Kumar et al. 2020] approached the problem of ranking clarification questions as a natural language inference task, aiming to determine if a question pertains to missing information. The development of datasets specifically designed for clarification question generation, such as [Aliannejadi et al. 2020], has been instrumental in advancing research in this area.

Dialogue Systems heavily rely on CQs to handle user utterances that are ambiguous, incomplete, or misunderstood. Recent work explores how dialogue systems can assess their own uncertainty to determine when and what type of clarification is needed. For instance, [**?**] investigated using model uncertainty to guide the generation of clarification questions, leading to significant improvements in task success. This aligns with the broader goal of making human-computer interactions more natural and efficient, moving to more targeted and context-aware clarifications [Purver et al. 2003]. The integration of large language models (LLMs) has further opened new avenues for generating more sophisticated and contextually relevant CQs in conversational AI, although challenges remain in ensuring their relevance and avoiding irrelevant questions in complex tasks [Google AI 2024].

The intersection of these domains highlights a shared goal: enabling systems to proactively identify ambiguity or missing information and engage users in a clarifying dialogue towards a decision. While significant progress has been made in generating and ranking CQs, ongoing challenges include discerning the optimal moment to ask a CQ, predicting its usefulness, and adapting CQ generation to diverse domains and user needs. Furthermore, the generation of CQs based on a contextualized knowledge base, extracted from organizational heterogenous data sources, is a new approach in decision support systems.

5. Methodology

To address the challenges of knowledge exploration in decision-making processes, we propose KnowOntoAsk, a knowledge-driven framework designed to guide individuals through goal-oriented, context-aware interactions supported by formal knowledge representations. The core premise of KnowOntoAsk is that meaningful questions—and consequently, better decisions—can emerge when individuals are supported not only in formulating their queries, but also in contextualizing them based on organizational knowledge and past experiences. The framework introduces a conceptual architecture that brings together three pillars: (i) knowledge bases derived from heterogeneous data sources, (ii) a dialogue-based exploration process grounded in formal ontologies, and (iii) support mechanisms for onboarding, clarification questions, and organizational memory. The architecture is designed to transform dispersed and unstructured organizational knowledge into reusable, navigable assets that drive human-AI collaboration in decision-making. The methodology includes the following Dimensions:

- *Development of a method to derive formal ontologies* from heterogeneous data sources, enabling semantic unification across structured and unstructured content.
- *Integration of generative AI models* (e.g., ChatGPT, Gemini) capable of engaging in natural language dialogues grounded in the generated ontologies, rather than relying solely on pretrained web-scale knowledge.
- *Implementation of onboarding and clarification mechanisms* that help individuals articulate their initial objectives and iteratively refine their questions during the exploration process.
- *Adoption of an Agentic AI architecture*, composed of specialized agents (e.g., clarifier, ontology builder, retriever, memory tracker) coordinated to execute dialogue, reasoning, and retrieval tasks.

- *Evaluation of the framework* through two real-world case studies: one in the academic domain, supporting research planning and knowledge reuse, and another in the financial sector, focused on contract analysis and risk assessment.

The core data flow, as illustrated in Figure 1, begins with the user initiating an interaction with an Intent (1), representing their information need or decision-making goal. This intent is then transformed into a Question (2) by the user, that goes over an Onboarding phase (3), which serves as the initial input to the system. The user's question is subsequently converted into a structured query (4), suitable for retrieving information from relevant knowledge sources. This query is then executed against an External Knowledge Base (5), which represents a broad repository of information relevant to the domain, generated earlier from heterogenous data sources. Initial information based on the query is retrieved from the External Knowledge Base, feeding it into a local knowledge, called Local Memory (6). The local knowledge acts as a contextual workspace, storing information relevant to the current dialogue and user intent.

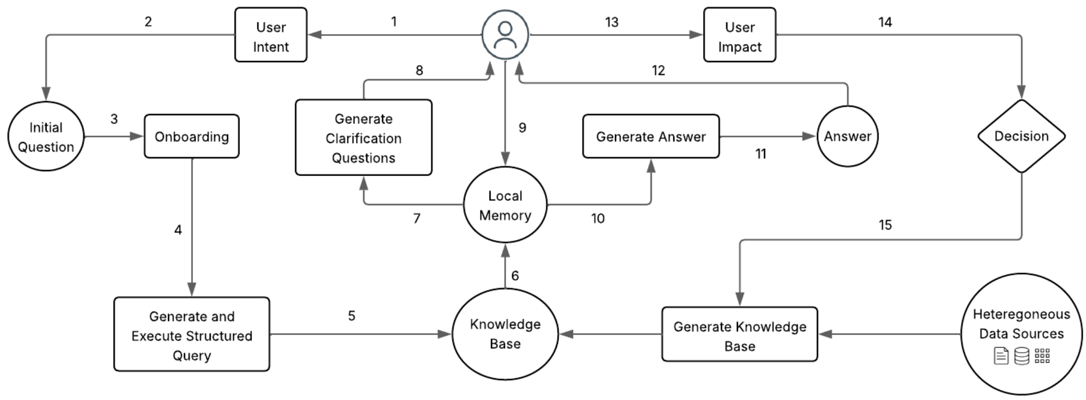


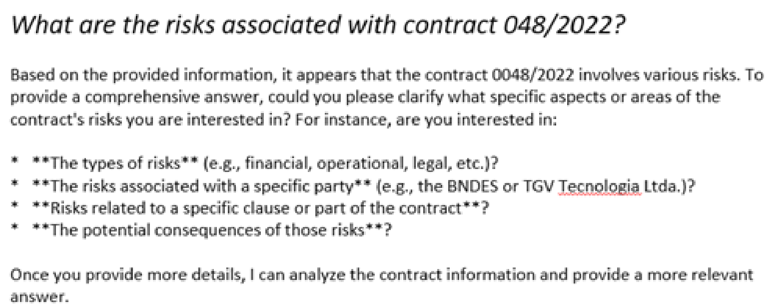
Figure 1. From Intent to Impact in Knowledge Exploration.

One of the key components of the framework is the Clarifying Questions generation module (7). Based on the information retrieved from the External Knowledge Base and the context within the Local Memory, the system generates targeted questions to address potential ambiguities or gaps in the user's initial query. These questions are presented to the user (8), prompting them to provide additional details. The user's responses (9) are then incorporated into the Local Knowledge, enriching the contextual understanding of the dialogue.

Once all necessary clarifying questions have been answered, and the Local Knowledge is deemed sufficient (10), the system generates an Answer (11) based on the refined information. This answer is presented to the user (12), which may generate an Impact (13), or change the Intent (1). The impact potentially drives a Decision (14) that leads to a specific Action, which is added to the knowledge base (15). This iterative process ensures that the system's responses are contextually relevant and directly address the user's information needs.

6. Preliminary Results

In the domain of Public Contract Management, the documents were transformed into a formal representation using KnowOntoAsk's ontology-driven modeling pipeline. The resulting knowledge base was then integrated into a dialogue system, enabling interactive and context-aware exploration of contractual information.



What are the risks associated with contract 048/2022?

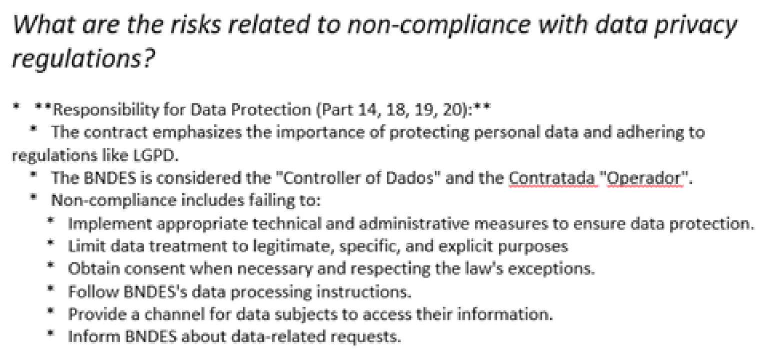
Based on the provided information, it appears that the contract 0048/2022 involves various risks. To provide a comprehensive answer, could you please clarify what specific aspects or areas of the contract's risks you are interested in? For instance, are you interested in:

* **The types of risks** (e.g., financial, operational, legal, etc.)?
* **The risks associated with a specific party** (e.g., the BNDES or TGV Tecnologia Ltda.)?
* **Risks related to a specific clause or part of the contract**?
* **The potential consequences of those risks**?

Once you provide more details, I can analyze the contract information and provide a more relevant answer.

Figure 2. Beginning risk assessment exploration of a particular contract.

For example, when prompted with a vague or high-level question—such as "What are the risks associated with contract 048/2022?"—the system does not attempt to provide a premature or overly generic answer. Instead, it initiates an interactive dialogue by asking targeted follow-up questions to uncover the user's intent, As the user progressively answers these clarification questions, the system is able to refine its understanding of the query and respond with context-rich, actionable insights derived from the formalized contract.



What are the risks related to non-compliance with data privacy regulations?

* **Responsibility for Data Protection (Part 14, 18, 19, 20):**
  * The contract emphasizes the importance of protecting personal data and adhering to regulations like LGPD.
  * The BNDES is considered the "Controller of Dados" and the Contratada "Operador".
  * Non-compliance includes failing to:
    * Implement appropriate technical and administrative measures to ensure data protection.
    * Limit data treatment to legitimate, specific, and explicit purposes
    * Obtain consent when necessary and respecting the law's exceptions.
    * Follow BNDES's data processing instructions.
    * Provide a channel for data subjects to access their information.
    * Inform BNDES about data-related requests.

Figure 3. Progressing on formulating better questions.

These preliminary results demonstrate the potential of KnowOntoAsk to support complex decision-making tasks through a dialogue-based approach that combines formal knowledge modeling, clarification strategies, and generative AI. This experiment provides initial validation that the framework can effectively bridge the gap between unstructured user queries and structured, semantically grounded answers.

7. References

References

Aliannejadi, M., Azzopardi, L., Kiesel, J., and Croft, W. B. (2024a). Clarifying the path to user satisfaction: An investigation into clarification usefulness. *arXiv preprint arXiv:2402.01934*.

Aliannejadi, M., Kiesel, J., and Croft, W. B. (2024b). Responses to conversational information retrieval clarifying questions with user emotions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Accepted to EACL 2024. Also on arXiv:2402.01934.

Aliannejadi, M., Kiesel, J., Zamani, H., and Croft, W. B. (2020). Generating clarifying questions for open-domain conversational search. In *Proceedings of The Web Conference 2020*, pages 1029–1040.

Dewey, J. (1910). How we think. *D.C. Heath & Co., Boston*.

Eberhart, Z. and McMillan, C. (2022). Generating clarifying questions for query refinement in source code search. *arXiv preprint arXiv:2201.09974*.

Ginzburg, J. (2001). Understanding and answering questions. *Language and Communication*, 21(4):331–360.

Google AI (2024). Learning to clarify: Multi-turn conversations with action-based contrastive self-training. `https://research.google/blog/learning-to-clarify-multi-turn-conversations-with-action-\based-contrastive-self-training/`. Accessed: June 7, 2025.

Kumar, V., Saha, P., and Mitra, C. (2020). Ranking clarification questions via natural language inference. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 3109–3112.

Purver, M., Ginzburg, J., and Healey, P. (2003). Clarification in spoken dialogue systems. In *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Dynamic Settings*, pages 56–63.

Rao, S. and Daumé III, H. (2018). Learning to ask clarifying questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1119–1129.

Zamani, H., Dumais, S. T., Craswell, N., Bennett, P. N., and Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 1018–1028.