poolparty®

WHITE PAPER

THE SEVEN CASES FOR KNOWLEDGE GRAPH INTEGRATION IN A RAG ARCHITECTURE

Andreas Blumauer – CEO of Semantic Web Company

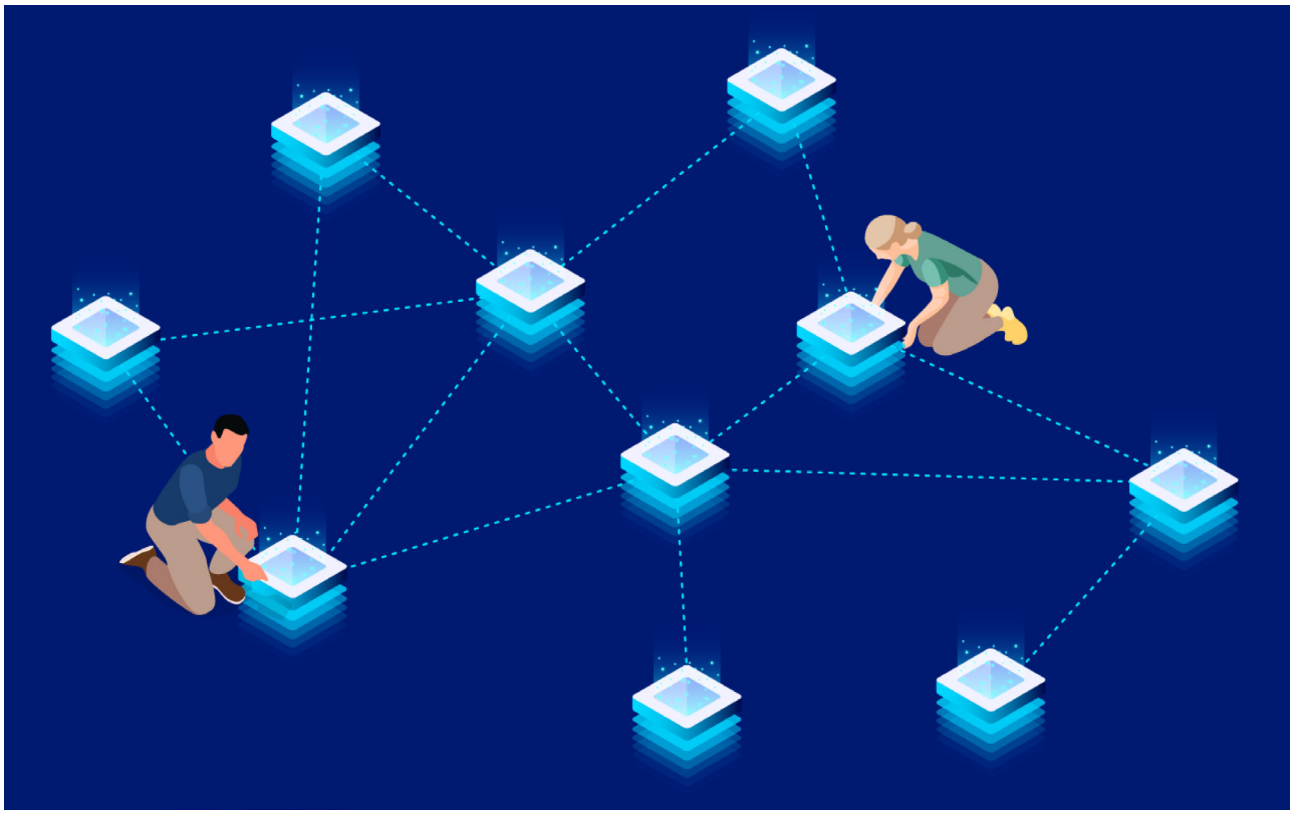


TABLE OF CONTENTS

INTRODUCTION 2

1. PROVIDE ADDITIONAL CONTEXT FROM KNOWLEDGE MODELS 3

2. PROVISION OF LINKED FACTS WITH THE HELP OF KNOWLEDGE GRAPHS 3

3. MAKE USE OF EXPLAINABLE REASONING 4

4. PERSONALIZATION 4

5. FUSE STRUCTURED CONTENT WITH KNOWLEDGE MODELS 4

6. EFFICIENT FILTERING OF RESULTS 5

7. USER QUERY ASSISTANT 5

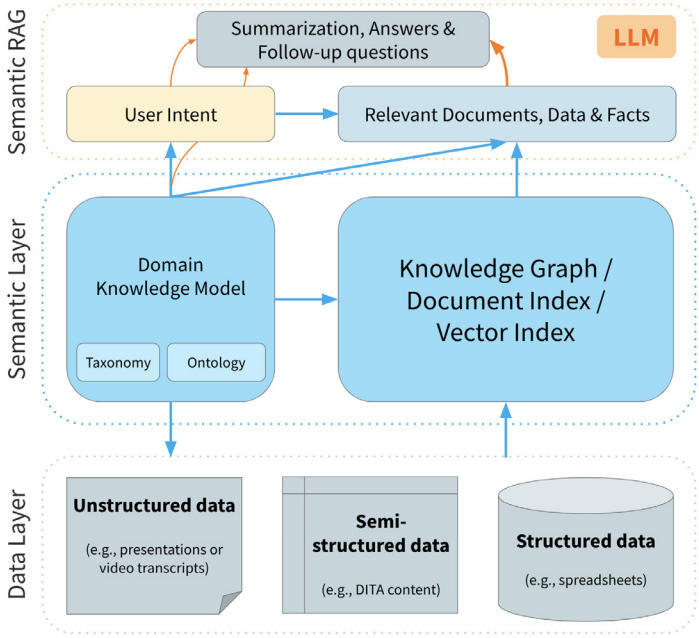
CONCLUSION 6

INTRODUCTION

In recent months, many companies have started to make use of LLMs and have implemented RAG architectures to support various business cases. A lot of dust has been kicked up, and it almost seemed as if even the most complicated information management challenges could finally be solved fully automatically with the help of generative AI. But after numerous experiments with real data, disillusionment set in: in many cases, the results were inadequate — the answers were unspecific or too specific, incomplete or simply wrong. The reasons for this ranged from insufficient data quality, contradicting information, data formatting issues over misunderstandings regarding the use of LLMs to the lack of substantial knowledge and content sources to ultimately find useful answers to tricky questions, especially in critical business processes.

Contemporary RAG architectures combine LLMs with vector databases for document search, but their use cases are typically limited to less critical processes. Often, only document-centric applications are covered (instead of integrating all relevant data across the enterprise), and RAG does not fully cover a user's typical search process.

Such user interaction should start with the formulation of an appropriate, well-specified search query (user intent) and end with convenient refinement methods such as search facets or content recommendations to make the search and serendipity process more efficient by providing concise answers or summaries and pre-formulating meaningful follow-up questions.



Recent trends now point to a fusion of symbolic AI such as knowledge models and graphs with statistical AI such as GenAI. RAG systems will no longer rely solely on vector databases, but also on domain knowledge models that provide additional contextual information about the respective knowledge area, and on graphs that enable efficient access to different, often very heterogeneous knowledge bases within an organization.

This white paper provides an overview of the benefits that can be achieved by a configuration that merges GenAI with Vector ***and*** Graph.

1. PROVIDE ADDITIONAL CONTEXT FROM KNOWLEDGE MODELS

Together with the enterprise knowledge graph, which serves as a universal data and content access point, domain knowledge models are the central building blocks of an organization's semantic layer. Knowledge models describe specific domain knowledge in a standardized way and help to organize and link business entities and concepts on a terminological, taxonomic, ontological and logical level. Knowledge models serve as a blueprint for the automatic creation of large enterprise knowledge graphs to link structured and unstructured data along user logic rather than inherent database logic and make it available via standards-based APIs. Such models contain knowledge about domains that would not be found anywhere else, neither in content systems nor in databases or spreadsheets, and thus also serve as a core element for enriching prompts with relevant information.

If someone asks: "How can CCOs keep up to date with new ESG regulations?", then an expert will be clear about what CCO stands for. In the context of ESG regulations, CCO most likely refers to the Chief ***Compliance*** Officer and not the Chief ***Communication*** Officer as one might assume without expert knowledge.

Semantic knowledge models can provide contextual information at various stages of a retrieval process, from query and user intent interpretation over model-driven prompt engineering and indexing, to personalization and filtering of results. In general, additional context helps RAG systems to better interpret and understand the relationships between entities and concepts found in questions and knowledge bases and their meaning for a particular user, thus improves the generalization capability of the model and reduces overfitting.

2. PROVISION OF LINKED FACTS WITH THE HELP OF KNOWLEDGE GRAPHS

Not every question can be answered most efficiently by combing through extensive document inventories. Rather, structured, fact-based databases provide the right answer, or should at least be consulted as one of the sources in a RAG system. When it comes to linking unstructured data (documents) with structured data (e.g. relational databases), vector databases do not offer a sufficient alternative to graph databases.

Using queries supported by vector databases, relevant parts of unstructured data can be efficiently addressed, but applying the same process to heterogeneous, including structured, data sets is more difficult with vectors because there is no explicit information about semantic relationships, neither at the instance level (taxonomic) nor at the schema level (ontological), but these are replaced by probabilistic models. Knowledge graphs are capable of ingesting both structured and unstructured information and are able to maintain the semantic understanding.

Example: You want your RAG to answer questions like: "Which spare part for machines that fulfill safety standards XYZ is on stock today?". Graph RAG is able to efficiently pool the required structured and unstructured data silos in a company and generate holistic views of data sets.

In addition, the use of knowledge graphs is also important as part of an AI strategy for curbing RAG proliferation: it can be assumed that more and more LLM-driven applications will be used in companies. As a result, an increasing amount of synthetic content and data will be available, which will subsequently be reused in RAG or fine-tuning steps. Provenance information is thus largely lost. The semantic layer and the knowledge graph

in a company's information architecture are therefore increasingly perceived as an immutable reference that ensures that no data smoothie has been created at the end of the AI hype which does not allow any conclusions to be drawn about essential metadata such as origin or quality.

3. MAKE USE OF EXPLAINABLE REASONING

The explainability and trustworthiness of AI results will apparently soon be required by law in the USA and is already a legal requirement in the EU. Furthermore, it should always be borne in mind that correlative models such as LLMs and vector databases only ever simulate causalities and can therefore in no way replace causal models. Automatic reasoning without being based on explicit models remains untraceable and therefore a black box approach.

Imagine you want to ask about “The most important non-financial risks for a company that are mitigated by the EU AI Act”, and your RAG system provides a list of risks that can be classified as more or less meaningful, but every time you ask the question just a little differently, a new list emerges. The follow-up question, “Why these and no other risks?” then also provides a wide array of explanations, depending on how the moon is currently positioned. This is not what we as risk managers like to see — we want comprehensible and consistent answers.

Consistency and explainability are the properties that semantic knowledge models bring to RAG architectures. Coupled with the traceability of where the data comes from, which is provided by knowledge graphs, a degree of trustworthiness of AI can be ensured, which is the only way to enable serious use in critical applications.

4. PERSONALIZATION

In a business context, even the best tuned LLM and the very best search algorithm will fail if the user's intent cannot be interpreted correctly. As explained in section 1, this starts with the resolution of ambiguities and continues with the correct interpretation of a question depending on the role and task spectrum, i.e. the individual information needs of the respective user. In short, what a Q&A engine urgently needs is personalization.

The CFO, the CIO and the head of marketing will each need very different answers and reference documents to the question "What does the EU AI act mean for our company?" in order to perform their tasks efficiently.

In a risk knowledge model, for example, it is possible to explicitly trace and define the key topics for different roles and areas. This contextual information helps to direct the focus of the RAG system to key aspects and allows to display material information in each case, e.g. to display essential safety information at the top depending on the user's range of tasks.

5. FUSE STRUCTURED CONTENT WITH KNOWLEDGE MODELS

There is a broad consensus that there is structured and unstructured data in companies. These two data pools are often treated completely separately from each other, not only technically but also organizationally, which always leads to silos and inconsistencies. As if this were not problematic enough, there are often other realities between these two worlds, which are referred to as semi-structured data, such as CSV files or structured (componentized) content.

If you want to develop a high-quality helpdesk that offers reliable support for all questions relating to a product or service, and you already have structured content for this, which may also be based on standards such as DITA, then you almost have the "golden bullet" in your hand to feed a corresponding RAG system. By combining a knowledge model with structured content, a knowledge graph is created that is also "open" for linking to other data sources. This knowledge graph as a content source for a semantic RAG system leads to more precise and specific answers, which can now also be traced back to micro-content.

6. EFFICIENT FILTERING OF RESULTS

The precise and efficient search for documents and sections within them will remain the central feature of most retrieval systems in a corporate context. Although synthetically generated summaries and quick answers or overview texts are gaining in importance, knowledge workers must still be able to access the original knowledge sources themselves, especially in regulated industries.

Traditional search tools such as faceted search are therefore by no means becoming obsolete, but remain an important component of a comprehensive enterprise information retrieval system. Vector-based RAG systems alone cannot solve this aspect, as they are imprecise and less performant, especially when it comes to the exact filtering of individual data and content objects along defined metadata and classification systems.

Semantic metadata and knowledge models form the basis for precise classification and faceting, but also for content recommendation systems that do not simply suggest "more of the same". If you also want complementary content to be displayed and filtered out, the retrieval system must access knowledge models where semantic relationships, not just correlations between metadata, are defined.

7. USER QUERY ASSISTANT

Simple RAG architectures are also inadequate for another reason: they do not take into account the entire workflows involved in a question-and-answer process and offer no way of making the formulation of initial inquiries and follow-up questions more efficient. In many industries, there is a shortage of skilled workers, which leads to enormous efforts to bring new talent on board. However, this is precisely where RAG-based dialog systems could provide significant support, but only if newer employees in particular are supported in their learning process. The key to a successful dialog, namely the formulation of questions, can only be supported in a precise manner with semantic knowledge models.

Here is an example: The question "Which diet can be used to support the treatment of diabetes?" can be made more precise with the help of a model-supported question formulation assistant. The assistant would ask whether the type of diabetes can be specified, or if other characteristics (e.g. age or secondary diseases) or the patient's life situation (e.g. pregnancy) that are decisive for answering the question are known.

An explicit domain model ensures that the contextualization of the question is carried out to an exhaustive and comprehensible extent, which can also correspond to a defined specification, instead of using LLMs to generate further contextual information rather by chance.

CONCLUSION

This paper discusses the limitations of vector-based RAG architectures in providing accurate and specific answers to complex questions. It highlights the importance of integrating knowledge graphs and symbolic AI (such as knowledge models) with statistical AI (such as LLMs) to improve the performance of RAG systems. The practical benefits to grounding an LLM in a knowledge graph are:

IMPROVED CONTEXTUAL UNDERSTANDING: Knowledge models provide additional precise contextual information about the knowledge domain, enabling better interpretation of user intent and more accurate responses.

PROVIDE LINKED FACTS: Knowledge graphs can store and process both structured and unstructured data on the basis of a consistent semantic knowledge model, enabling efficient and uniform access to different knowledge bases within an organization.

EXPLAINABLE REASONING: Semantic knowledge models ensure the explainability and trustworthiness of AI results, which is critical for mission-critical applications.

PERSONALIZATION: Knowledge models can be used to personalize responses based on user roles and tasks, ensuring that relevant information is displayed for each user.

FUSE STRUCTURED CONTENT WITH KNOWLEDGE MODELS: Combining structured content with knowledge models creates a knowledge graph that can be used to enhance a RAG system to provide accurate and specific answers.

EFFICIENTLY FILTER RESULTS: Semantic metadata and knowledge models enable precise classification and faceting, as well as content recommendation systems that suggest relevant content from diverse sources.

USER QUERY ASSISTANT: Semantic knowledge models can help formulate initial queries and follow-up questions, making the question-and-answer process more efficient.

In summary, RAG systems should not rely solely on vector databases, but should integrate knowledge models and graphs to provide comprehensive and reliable answers to complex questions. With Semantic RAG one can keep domain fidelity and response focus high and hallucinations low. Often referred to as “Advanced RAG” in the literature, Semantic RAG (or Graph RAG) is a cascade of contextual methods and LLM calls to improve all processes in an enterprise retrieval system from analyzing user intent to filtering results quickly, efficiently and in an explainable manner.

POOLPARTY SOFTWARE / SEMANTIC WEB COMPANY

Semantic Web Company GmbH (SWC)
Mariahilferstrasse 70 / Neubaugasse 1, Top 8
A-1070 Vienna, Austria
Phone: +43 1 402 12 35
email: office@semantic-web.at

Semantic Web Company (SWC) is the leading provider of graph-based knowledge technologies. Its flagship product, the PoolParty Semantic Suite, is the most complete semantic middleware platform on the global market comprising taxonomy, auto classification, language processing, text mining, semantic search, and knowledge graph solutions.

For several years in succession, the company has been featured in Gartner publications as a "Sample Vendor" in the Gartner Hype Cycles and a "Visionary" in the Magic Quadrant for Metadata Solutions. Additionally, SWC is proud to be continuously named a "Company that matters in knowledge management" and a "Trailblazer" for Semantic AI in KMWorld's annual reports.

With a customer base of Global 2000 companies, SWC is dedicated to helping customers roll out their knowledge management and AI strategies within their organizations.

Semantic Web Company is headquartered in Austria and has branches in the US, UK and France. To learn more, visit www.semantic-web.com and www.poolparty.biz or follow them on LinkedIn and Twitter.