



**Case study: Use of the meaning-based search engine  
SEMPRIA-Search as enterprise search in a consulting  
company**

**SEMPRIA GmbH  
Grafenberger Allee 277–287  
40237 Düsseldorf, Germany  
<https://www.sempria.de/>  
[info@sempria.de](mailto:info@sempria.de)**

Last revision: 2025-11-17



# Contents

<b>1</b>	<b>Preliminary remarks</b>	<b>2</b>
1.1	Definition of enterprise search . . . . .	2
1.2	Use of AI . . . . .	2
<b>2</b>	<b>Challenges and requirements for an enterprise search</b>	<b>2</b>
2.1	Variety of sources . . . . .	2
2.2	Formats: of all kinds and from all eras . . . . .	3
2.3	Metadata diversity . . . . .	3
2.4	Text documents: even without text? . . . . .	4
2.5	Multimedia: beautiful, but without words . . . . .	4
2.6	Graphical User Interface (GUI) . . . . .	4
2.7	Reading rights and access rights: who is permitted to do what? . . . . .	5
2.8	Babylon: multilingual instead of monolingual . . . . .	5
2.9	Duplicates: ... and another copy ... and again as a PDF . . . . .	6
2.10	Data protection and GDPR . . . . .	6
2.11	User-defined document collections . . . . .	6
2.12	Domain knowledge . . . . .	6
2.13	External data sources . . . . .	7
2.14	Integrated text technology features . . . . .	7
<b>3</b>	<b>Improving the quality of search results through cognitive meaning-based search</b>	<b>8</b>
<b>4</b>	<b>Evaluation</b>	<b>9</b>
<b>5</b>	<b>Next steps</b>	<b>9</b>
		1

# 1 Preliminary remarks

This case study analyzed the design, implementation, and ongoing evaluation of the meaning-based search engine SEMPRIA-Search in a medium-sized consulting company with over 100 employees. For simplicity, this company will be referred to as **IT42consult** in the following text. The case study aimed to investigate the advantages and potential of a deep semantic (i.e., meaning-based) search as an enterprise search engine.

## 1.1 Definition of enterprise search

For an enterprise search, all relevant documents of a company or organization should be searchable as full texts. Important metadata such as creation date, title, areas, access rights etc. should be included. Document sources include file systems, Microsoft SharePoints, mail servers, content management systems (CMS), document management systems (DMS), wikis, and external websites.

## 1.2 Use of AI

We have been employing artificial intelligence (AI) methods for over 20 years. Our AI components prove their worth by using human-readable representations and linguistic concepts at their core. This approach is known as symbolic AI (or conceptual AI or representational AI). Symbolic AI differs significantly from generative AI (GenAI) in its advantages and disadvantages. Results and inferences are always verifiable and explainable; hallucinations, misinformation, and lies, as seen with LLMs, can thus be avoided. On the other hand, symbolic AI currently performs only a few inference steps (e.g. how *export* and *import* are semantically related), which makes a big difference for a better search performance. GenAI can be integrated into the search solution upon customer request, for example, through retrieval-augmented generation (RAG). For enterprise search, a solid clever search engine is often the best choice.

# 2 Challenges and requirements for an enterprise search

## 2.1 Variety of sources

In companies, you will usually encounter a mixture of the following document sources:

- Microsoft SharePoint, also multiple SharePoints
- Intranet
- File systems
- Emails



- Content management systems (CMS)
- Document management systems (DMS)
- Wikis and
- Websites (own and third-party)

In the company of the case study, the following sources were identified in several rounds of discussions with the technical and specialist contacts:

- 2 large SharePoints with approximately 200,000 documents
- several Windows Server file systems with approximately 400,000 documents
- 3 external websites (from national and international industry organizations or regulators) with approximately 6,000 documents

## 2.2 Formats: of all kinds and from all eras

The formats change along with the sources. Office formats like DOCX, PPTX, and XLSX appear, but also older gems like RTF. And of course, countless PDFs. For each format, the precise text extraction must be programmed. This is best done in logical reading order, which is not easy with many PDFs due to multiple columns and footnotes.

## 2.3 Metadata diversity

Website search can often import a comprehensive set of metadata (author, date written, date modified, keywords, etc.) from the CMS. For enterprise search, a transfer to a harmonized metadata schema of the search engine must be programmed for each source system. And sometimes it is more accurate or reliable to extract metadata such as the date written from the name or beginning of the document.

For **IT42consult**, the geographical restriction of the search (e.g. by the client's federal state) was a requirement that arose from feedback received during the first 12 months. To achieve this, an existing company database was appropriately exported so that each project, and therefore each document, was assigned a location (and thus a state).

The metadata allows a powerful restriction of search results using faceted search. After an initial search, it is shown that the results can be restricted, for example, by year (2024, 2025) and document type (PDF, XLSX).

In discussions with users in the case study, the most important facets were selected and arranged in order of importance in the search mask. To avoid overloading the interface, some special cases were hidden. These decisions were recorded for annual review.

Search engine results pages (SERPs) should present the metadata of the results in a readable and clear manner. For **IT42consult**, a few metadata attributes were hidden to



ensure that a sufficient number of search results could be displayed on one monitor page. Further details can be displayed, for example, by moving the mouse pointer (mouse-over) to certain areas.

## 2.4 Text documents: even without text?

The following disillusionment often arises after launching an internal search engine: *We have this great PDF that contains everything, but it can never be found.* It quickly becomes apparent that many PDFs (and other file formats) cannot provide any text for the search engine because they are only raster graphics from a scanner or camera. A good cognitive enterprise search engine uses optical character recognition (OCR) to make these documents accessible for analysis. Sometimes it is very worthwhile to train the OCR for the specific characteristics of the documents.

At **IT42consult**, it turned out that a considerable portion of the scanned PDFs already contained OCR results (from over 20 different programs). An evaluation showed that in the majority of cases, a current OCR program delivers significantly better results. Therefore, a new OCR result is produced for each scanned PDF with an included OCR result, and the best of the two is indexed for the search.

## 2.5 Multimedia: beautiful, but without words

Video and audio files often contain important information, but usually without text. Similar to OCR for scans, automatic speech recognition (ASR) can be used to add text that can be found via a text search.

In the document world of **IT42consult**, audio and video files play no role, so they were excluded from indexing.

## 2.6 Graphical User Interface (GUI)

When searching a website, a minimal search box of 20 characters is often sufficient. The situation is quite different with enterprise search engines. Here, there is much more metadata and information about the sources and source systems. An entire screen can quickly be filled with a search mask (see Figure 1), possibly with submasks. While this is a powerful tool for advanced users, it is more appropriate for the average searcher to also offer a simplified search interface that includes only the most important metadata and facet selections.



The image shows a search mask interface with the following elements:

- Area**: A label for the search area.
- Search**: A search button and a search input field containing the text "applications by Bayer".
- Language of query**: A dropdown menu set to "automatically determined".
- Language of hits**: A dropdown menu set to "monolingual (matching the query)".
- Interpretation**: A dropdown menu set to "semantic (recommended)".
- From (incl.)**: A dropdown menu set to "unlimited".
- Till (incl.)**: A dropdown menu set to "unlimited".
- Format**: A dropdown menu with options: "unlimited", "article without images", and "article with images".
- URL restriction**: An empty text input field.
- Title**: An empty text input field.
- Search in results**: An empty text input field.
- Order**: A dropdown menu set to "best first".
- Empty search form**: A button with a plus icon and the text "Empty search form".
- Home icon**: A small house icon at the bottom left.

Figure 1: Search mask as part of the user interface.

## 2.7 Reading rights and access rights: who is permitted to do what?

Many sources that feed an enterprise search have a sophisticated rights system (ActiveDirectory, LDAP, etc.). From a search engine perspective, it is particularly important to know who is allowed to read which documents. This must be accurately replicated by the enterprise search.

ActiveDirectory is used extensively at **IT42consult**, with hundreds of groups (some of which are technically required). The same applies to SharePoints. The rights from the source systems are projected onto the document rights of the indexed documents.

## 2.8 Babylon: multilingual instead of monolingual

Many website search engines are monolingual because only German documents are available. In the enterprise environment, multilingualism is the standard. German and English documents in particular should be found, regardless of whether search queries are formulated in English or German.

At **IT42consult**, automatic language identification for all documents showed that German accounts for about 80 % and English for about 20 %. Multilingual search is therefore highly justified here.



## 2.9 Duplicates: ... and another copy ... and again as a PDF

Duplicates are relatively rare in website search engines; in enterprise search engines, however, we see duplicate rates of 20 % to 75 %. Without duplicate detection, a duplicate rate of 75 %, for example, means that on average 3 out of 4 documents in the results are redundant. This can unnecessarily quadruple reading and research time!

Duplicates are not just byte-identical documents; they should also be detected across format boundaries, e.g. when a PDF was generated from a DOCX or XLSX. Which format is prioritized in the search results is part of the configuration or personalization of the internal search engine.

For the **IT42consult**, the duplicate rate is 24 %. Another particular challenge was that the search engine should only display one document per duplicate group as a result, while simultaneously making all duplicates readily available. This is particularly important if access rights in a duplicate group were not assigned consistently by hand; a case that was more common than anyone expected.

## 2.10 Data protection and GDPR

Company-wide search engines are often offered as cloud solutions hosted abroad (e.g. in the USA). Other options include operation in a foreign data center, a domestic data center, or on-premises as a server solution within the company itself.

Data protection plays a major role for **IT42consult**. In principle, no data should leave the company, so the server solution (on-premises) was ideal. Since the search engine does not use any third-party software or services, the data is just as well protected as the company's existing data. Technically, the search engine runs on a Linux virtual machine on an existing, underutilized server. This way the hardware costs could be avoided. Secure remote maintenance was implemented using state-of-the-art technology.

## 2.11 User-defined document collections

Many users want to define their own document collections, for example, by simply assigning a so-called tag to the documents on a search engine results page (SERP). These collections can then be used as a restriction in the search.

Document collections were a basic requirement for **IT42consult**. It was also important to be able to distinguish between private and public tags. A public tag is visible to the search engine's users and can be maintained cooperatively.

## 2.12 Domain knowledge

For many companies, it is worthwhile to integrate existing formalized knowledge (such as thesauri) into the search.



For **IT42consult**, this applied particularly for the complex part-whole relationships between smaller companies and the parent company, and for official, unofficial, and abbreviated spellings of company and institution names. There are still many specific abbreviations in the industry, some of which stand for different terms than in other industries. For example, *DB* can stand for database, decibel, a major bank, a railway company, and more. This domain knowledge was integrated into the enterprise search engine with little effort.

## 2.13 External data sources

The integration of Wikipedia (Wikidata, Wikivoyage, etc.) is a worthwhile endeavor for some institutions, provided that it can effectively supplement the institution's own document base.

No significant advantages were anticipated for **IT42consult** with its highly specialized domain, so no external data sources (as successfully implemented for other companies) were indexed.

## 2.14 Integrated text technology features

Once you have invested a lot of effort in setting up a company-wide search to index as many of an organization's documents as possible, it makes sense to consider what else you can do with this treasure trove of documents. Some popular text technology functions can be mentioned as ideas and building blocks:

- automatic subject indexing and keywording, facets
- readability assessment for texts
- detection of repetitions and plagiarism (at the semantic level, not just at the character level)
- finding contradictions
- info boxes, link boxes, topic pages and specials
- automatic summarization (abstracting)
- semantic version analysis of texts
- review of terminology (organization-wide spelling and technical terms)



### 3 Improving the quality of search results through cognitive meaning-based search

SEMPRIA-Search follows a cognitive, meaning-based approach, that means all documents and all search queries are translated into formal meaning representations. These can be transformed into representations for synonymous or similar formulations; inferences also work in this way. The advantage of this representational AI is that all results can be explained. Through this representational language understanding, the search engine brings together many different (but synonymous) formulations between the search query and the document text (resulting in greater search recall). On the other hand, it avoids false matches (resulting in greater search precision).

The **IT42consult** initially decided that no explanations should be displayed in order to avoid overloading the search engine interface. A review was scheduled for the agreed annual status meetings.

To give an impression of the nature and diversity of the linguistic phenomena dealt with in terms of synonymy, similarity, and inference, here are some important classes of phenomena.

- Expanded search with synonyms (*policeman - cop*)
- Expanded search using concept hierarchies, i.e., connections between subconcepts and superconcepts (*Neodymium is a rare earth*)
- Use of name variants (*Venetia - Veneto*)
- Expanded search with alternative formulations (*oversight - oversee, supervise*)
- Use of geographical knowledge (*Amberg is located in Bavaria*)
- Handling of ambiguity (*The Jaguar drove into a decrepit jaguar.*)
- Comparison of meanings, no comparison of character strings (*The BMI reports on the rising BMI.*)
- Co-references in texts (*Mrs. Miller - the actress - she*)
- Multilingual internal search engine (*armchair - Sessel; long-distance traffic - Fernverkehr*)
- Question function (*Who exports uranium?*)
- Various orthographies (*authorize - authorise*)
- Robustness against misspellings and spelling variants (*terace - terrace; footbal - football*)



## 4 Evaluation

Since the search engine installation in the case study is still quite new, not all evaluation data is available yet.

The success and increased quality compared to the previous solution (standard Share-Point search) can best be seen in the number of users and the number of search queries. With the number of employees at the company remaining constant, the number of active users gradually increased from 10 % to approximately 50 %. This is a very good percentage, as the workforce also includes external staff who, for technical and organizational reasons, are not supposed to use the company-wide search at all. The **IT42consult** occasionally offered a half-hour workshop on the topic of *in-house research* to promote the successful use of the new search engine. Search numbers rose even more significantly than user numbers, increasing by a factor of 8.

An important factor in the increase in both of these figures was that feedback (such as suggestions for new features, special requests, error reports) was collected in the **IT42consult** and the search engine manufacturer was able to offer improvements of the most important feedback topics to a test user group within the company as quickly as possible. If successful, these adjustments were then rolled out company-wide.

## 5 Next steps

The authors of this case study can be reached for questions at [info@sempria.de](mailto:info@sempria.de) and by phone at +49 211 566693-57. One-time and ongoing costs can be reliably calculated after determining just a few parameters.

Test systems are possible and often recommended. The costs of a test system (without special requests or major adjustments) are not charged to the interested party. As with the **IT42consult**, a representative sample of 2 % to 10 % of the document inventory is recommended as the basis for a test system.

