Bridging the Vocabulary Gap: Mapping Free-Text Queries to Controlled Thesaurus Descriptors Using Large Language Models

Arben Hajra[1 [0000−0001−5081−5927]] and Tamara Pianos[1 [0000−0002−4386−519X]]

ZBW – Leibniz Information Centre for Economics, Germany
{a.hajra,t.pianos}@zbw.eu

Abstract. Users of academic digital libraries increasingly express information needs through natural-language or free-text queries, while retrieval in digital library systems continues to rely on controlled vocabulary indexing. This vocabulary mismatch is particularly pronounced in domain-specific portals such as EconBiz, which uses the STW Thesaurus for Economics. Analysis of query logs shows that longer and natural-language queries are more frequently associated with zero-result retrieval, highlighting limitations of keyword-based interfaces for complex information needs. We present a multi-stage LLM-based retrieval pipeline for mapping free-text queries to controlled thesaurus descriptors. The pipeline combines AI-assisted concept extraction, hybrid lexical-dense retrieval over a vector-indexed thesaurus, cross-encoder reranking for primary-subject identification, and SKOS-aware hierarchical pruning. Evaluation on a title-as-query benchmark derived from EconBiz records, using expert-assigned STW descriptors as ground truth, shows consistent improvements in precision and recall over lexical and dense-retrieval baselines while reducing descriptor set size for Boolean retrieval. A complementary manual evaluation on real-world zero-result queries indicates that the approach can recover relevant retrieval paths for previously unsuccessful searches. The system is deployed as an internal interactive evaluation interface. The results suggest that LLM-based query-to-descriptor mapping improves natural-language access to specialized digital libraries while remaining compatible with established indexing infrastructures.

Keywords: digital libraries · controlled vocabularies · SKOS · thesaurus mapping · hybrid information retrieval · large language models · query expansion

1 Introduction

Controlled vocabularies, thesauri, and ontologies form a core layer of semantic interoperability in digital libraries and scholarly information systems [21,5,16]. Descriptor assignment, i.e., the association of documents with concepts from a controlled vocabulary, supports semantic retrieval, faceted navigation, and metadata harmonisation across heterogeneous collections. Standards such as

the Simple Knowledge Organization System (SKOS) [28] provide formal mechanisms for representing concepts, preferred labels, alternative labels, and semantic relations. While considerable work has addressed document-side descriptor assignment [15,23], query-side mapping to controlled vocabulary concepts has been examined in selected retrieval domains and in broader query expansion research [1,9,3,7], but has received comparatively limited attention in operational digital library systems and domain-specific retrieval infrastructures.

The broader information retrieval literature has long identified vocabulary mismatch as a central source of retrieval failure, particularly when user terminology diverges from indexing language [14,46,7]. Conventional keyword-based retrieval relies heavily on lexical overlap and often performs poorly when users formulate exploratory information needs using conversational language, paraphrases, or domain-adjacent terminology. Recent advances in large language models (LLMs) and dense retrieval provide improved contextual matching capabilities, but typically lack explicit grounding in controlled vocabularies, which can result in semantically plausible but structurally misaligned or overly general interpretations [51,38].

This challenge is also evident in EconBiz [11], the academic search portal operated by ZBW – Leibniz Information Centre for Economics [48]. EconBiz provides access to more than 13 million records indexed using the STW Thesaurus for Economics [43]. Extending prior analyses of EconBiz search behaviour by Zerhoudi and Granitzer [49], we analysed 30,984 user queries collected during April 2026. The resulting patterns are consistent with their findings and further substantiate similar issues in query complexity and user interaction behaviour within EconBiz. In particular, long and semantically complex natural-language queries such as "I am interested in the impact of AI on the European labour market after 2020"—frequently lead to sparse retrieval outcomes and repeated query reformulation. This behaviour reflects a persistent mismatch between natural-language search expressions and descriptor-oriented retrieval infrastructures.

To address this problem, we present a multi-stage query-to-descriptor mapping pipeline that converts natural-language economics queries into STW descriptors and integrates with EconBiz's existing retrieval infrastructure.

This paper makes three contributions:

- A pipeline for mapping free-text queries to STW descriptors, combining LLM-based concept extraction, hybrid lexical-dense retrieval, adaptive score fusion, cross-encoder reranking, and SKOS-aware hierarchical pruning.
- A large-scale title-as-query benchmark over the EconBiz catalogue, using cataloguer-assigned STW descriptors as ground truth, with evaluation against lexical and dense retrieval baselines.
- A manual evaluation on real-world zero-result queries, demonstrating the system's ability to recover meaningful retrieval paths under retrieval failure conditions.

The remainder of the paper covers related work (Sect. 2), resources and datasets (Sect. 3), system architecture (Sect. 4), evaluation (Sect. 5), discussion and limitations (Sect. 6), and conclusions (Sect. 7).

2 Related Work

2.1 Controlled Vocabularies and Semantic Indexing

Controlled vocabularies and ontologies are a foundational component of semantic interoperability in digital libraries and scholarly information systems [17,33]. Standards such as ISO 25964 and SKOS define formal models for representing concepts, labels, and semantic relations, enabling consistent indexing and cross-system interoperability [21,28]. Domain-specific thesauri, including STW, AGROVOC, and the OECD taxonomy, exemplify curated knowledge organisation systems designed to support structured subject indexing and retrieval.

Existing systems based on controlled vocabularies primarily focus on document-side descriptor assignment [15,23]. However, user queries are typically short, informal, and lexically divergent from controlled indexing language, creating a persistent gap between query expressions and curated descriptors [9,3]. Early work on thesaurus-based query expansion in bibliographic retrieval demonstrated that mapping user queries to controlled medical subject headings (MeSH) substantially improves recall in MEDLINE search [42], providing early evidence for query-side vocabulary bridging in domain-specific digital libraries.

2.2 Retrieval Models and Semantic Matching

Classical information retrieval methods rely on lexical matching, which is sensitive to vocabulary mismatch between queries and indexed documents. Early work has shown that such mismatch is a primary source of retrieval failure [39,14]. Embedding-based retrieval models address this limitation by learning distributed semantic representations that improve robustness to lexical variation [27,38]. Static word embeddings were shown early on to outperform classical vector-space retrieval on short texts such as titles [18], foreshadowing today's contextual LLM-based dense retrieval.

Modern systems frequently combine sparse lexical methods with dense semantic representations in hybrid architectures, improving both recall and ranking robustness [10,12,26]. Neural sparse retrieval models such as SPLADE [13] and ColBERT [24] extend this line through learned sparse representations and late interaction, integrating lexical and semantic signals without explicit vocabulary constraints. These models, however, operate over document or passage representations rather than curated descriptor spaces, so semantic relevance alone is insufficient – retrieved candidates must still align with structured indexing schemes, creating a persistent mismatch with generic relevance optimisation. Cross-encoder reranking is widely used as a second-stage approach, enabling fine-grained interaction modeling between queries and candidate documents [31,32,45].

Recent work in neuro-symbolic and knowledge-grounded retrieval explores integrating neural models with structured knowledge resources to improve interpretability and consistency [20,47]. These approaches preserve symbolic constraints while using neural components for interpretation or disambiguation, closely related to mapping queries to controlled descriptor representations.

2.3 Natural-Language Interfaces and Scholarly Search Systems

User interaction studies in digital libraries highlight persistent difficulties in query formulation and refinement. Zerhoudi and Granitzer [49] analyse EconBiz interaction logs and show that users often engage in extended search sessions with frequent reformulations, indicating challenges in expressing information needs in system-compatible form.

Recent scholarly search systems increasingly incorporate LLMs and retrieval-augmented generation (RAG) to support natural-language interaction [41,44,34]. These systems enable more flexible query formulation and synthesis of retrieved information, but typically require additional embedding-based indexing of document content and primarily operate at the level of document retrieval and answer generation [25] rather than controlled-vocabulary alignment.

Previous work investigated agentic RAG architectures integrated with knowledge graphs for scholarly exploration within EconBiz [19]. In contrast, the present work focuses on an earlier stage in the retrieval pipeline: mapping natural-language queries to controlled STW descriptors. Rather than replacing existing Solr-based retrieval infrastructures, the proposed approach augments them through structured query interpretation, enabling compatibility with established digital library indexing practices.

3 Resources and Data

3.1 STW Thesaurus for Economics

The STW Thesaurus for Economics, developed and maintained by ZBW, is a domain-specific controlled vocabulary for indexing literature in economics, business administration, finance, and related social sciences [43,30]. Published as Linked Open Data using the SKOS standard [28], STW comprises more than 6,000 standardised descriptors and over 20,000 entry terms.

Descriptors are organised through hierarchical relations (`skos:broader`, `skos:narrower`) and associative relations (`skos:related`), enabling structured navigation and concept expansion. The thesaurus is available in standard RDF serialisations and is continuously curated by domain experts to ensure conceptual consistency and topical coverage.

3.2 Descriptor Enrichment and Vector Indexing

To support semantic alignment between natural-language queries and controlled vocabulary descriptors, we enrich each STW descriptor with supplementary descriptive text. Where available, descriptions were derived from DBpedia abstracts [2] via existing `skos:exactMatch` alignments between STW concepts and DBpedia entities (3,877 mappings in total). For concepts without such alignments, concise auxiliary descriptions were generated using a locally deployed LLM with controlled prompting to ensure domain consistency. These descriptions are used solely as semantic augmentation and do not replace authoritative thesaurus definitions.

Each descriptor representation includes bilingual preferred labels (English and German), alternative labels, semantic relations, and descriptive text. To support retrieval, descriptor labels, alternative labels, and descriptive text were embedded in separate vector collections using two dense embedding models, BGE-M3 (567M parameters) [8] and Qwen3-Embedding (8B parameters) [50]. These models were selected due to their strong performance on the Massive Text Embedding Benchmark (MTEB) [29] and their multilingual capabilities, which are particularly relevant for the EconBiz collection. The remaining descriptor metadata, including semantic relations and auxiliary attributes, was stored as payload within each collection.

The resulting vectors were indexed in a Qdrant vector database [36], enabling efficient approximate nearest-neighbor retrieval for candidate descriptor generation and forming the semantic retrieval layer of the proposed system [22].

3.3 Evaluation Dataset

EconBiz is ZBW's portal for scholarly literature in business and economics, providing access to more than 13 million bibliographic records. A substantial subset of these records is indexed with STW descriptors, around 4.2 million records (about 32% of the collection) have at least one assigned STW descriptor.

For the title-as-query evaluation (Section 5.1), we sampled 1,000 records stratified across top-level STW categories to ensure broad thematic coverage. Records with fewer than two or more than ten descriptors were excluded to reduce under-specified and overly broad subject assignments.

For the manual evaluation (Section 5.2), we selected 50 user queries from EconBiz interaction logs that previously resulted in empty retrieval outcomes in the standard keyword-based interface.

3.4 Empirical Motivation from Query Logs

To establish the practical motivation for our system, we analysed a sample of anonymised EconBiz query logs. Our analysis focuses on the relationship between query formulation and retrieval failure, with particular attention to how increasing query length and natural-language complexity affect retrieval effectiveness. We analysed anonymised EconBiz query logs from April 2026 comprising 30,984 queries: 22,261 short queries ($\leq$ 4 words), 4,967 medium-length queries (5–8 words), and 3,756 long queries ($>$ 8 words). The analysis focuses on general free-text searches, excluding known navigational and field-restricted searches (e.g., title, author).

Following established practices in scholarly information retrieval and findings from the TREC Interactive Track literature [4], we define retrieval failure as a query returning fewer than five results. Although such queries are not strictly empty, sparse result sets do not provide sufficient coverage for exploratory literature search and are typically associated with query reformulation in interactive search sessions.

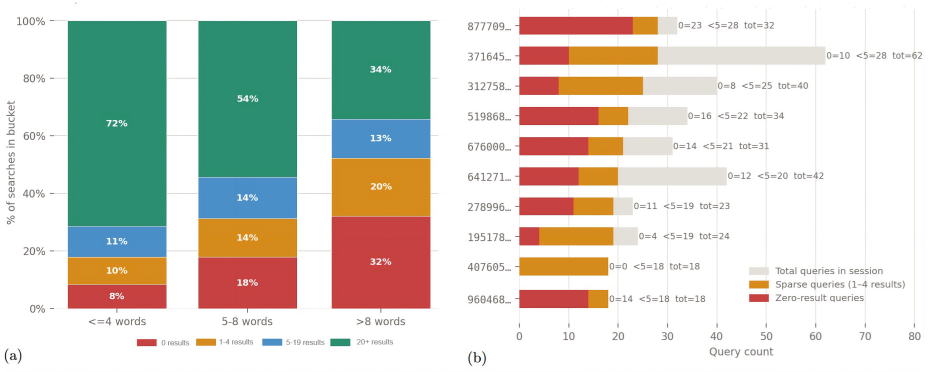

Fig. 1. (a) Result count distribution by query length in the EconBiz query log. (b) Top 10 sessions ranked by low-result queries per session. Data collected in April 2026.

Figure 1(a) shows the result count distribution stratified by query length. Short queries ($\leq$ 4 words) fail in 17.8% of cases; medium-length queries (5–8 words) remain comparable at 18.0%, though the proportion of zero-result queries increases. For long, naturally phrased queries ($>$ 8 words), the failure rate rises sharply to 52.3%: more than half return fewer than five results, and 32% return no results at all.

The contrast between long and short queries (52.3% versus 17.8%) constitutes the central finding of this analysis. It indicates that users expressing information needs as natural-language research questions experience substantially lower retrieval effectiveness than users issuing shorter keyword-oriented queries.

This effect is particularly significant because long queries, although the smallest group by volume, often reflect the most semantically rich and specific information needs. The dataset contains 3,139 unique zero-result queries, representing concrete vocabulary mismatches where user terminology fails to align with the controlled indexing vocabulary.

Figure 1(b) shows the top 10 sessions ranked by the number of low-result queries within a single session. The most extreme case involved 32 queries, 28 of which returned fewer than five results, including 23 with zero results, consistent with repeated reformulation attempts during systematic literature searching. Across the full dataset, 8.8% of sessions contained three or more consecutive low-result queries, a behavioural pattern indicative of repeated unsuccessful reformulation.

Taken together, these findings indicate that retrieval failure is not random but systematically associated with increasing query complexity and natural-language formulation. This pattern reveals a structural mismatch between users' expressive search behaviour and the descriptor-oriented indexing vocabulary underlying EconBiz.

The query-to-descriptor mapping pipeline introduced in Section 4 is designed specifically to address this mismatch by translating natural-language information

needs into formally grounded STW descriptors compatible with the existing retrieval infrastructure.

3.5 System Infrastructure

The system is deployed on institutional GPU infrastructure with local model execution. All language models are run using Ollama [35], ensuring fully local inference without external API dependencies and compliance with data protection requirements. The retrieval and processing pipeline is containerised using Docker to ensure modularity and integration with existing EconBiz services.

4 System Architecture

The pipeline addresses the query-side vocabulary gap in controlled-vocabulary retrieval by mapping free-text queries onto formally grounded STW descriptors, which are subsequently translated into structured Boolean queries over the EconBiz Solr index. The architecture is organised into four functional phases: Query Understanding, Descriptor Retrieval, Descriptor Refinement, and Interactive Query Construction, as illustrated in Figure 2.

The system is designed around three operational constraints: (i) strict grounding in the STW controlled vocabulary to ensure verifiable descriptor assignments against published SKOS metadata, (ii) compatibility with the existing EconBiz Solr infrastructure without requiring index or schema modifications, and (iii) low-latency interaction suitable for real-time search assistance. To satisfy these constraints jointly, the pipeline combines lexical retrieval, dense semantic matching, LLM-based concept extraction, neural reranking, and symbolic SKOS-aware post-processing in a staged hybrid architecture.

4.1 Phase 1: Query Understanding

LLM-based concept extraction. The query understanding stage employs a locally deployed LLM to extract structured query representations for subsequent

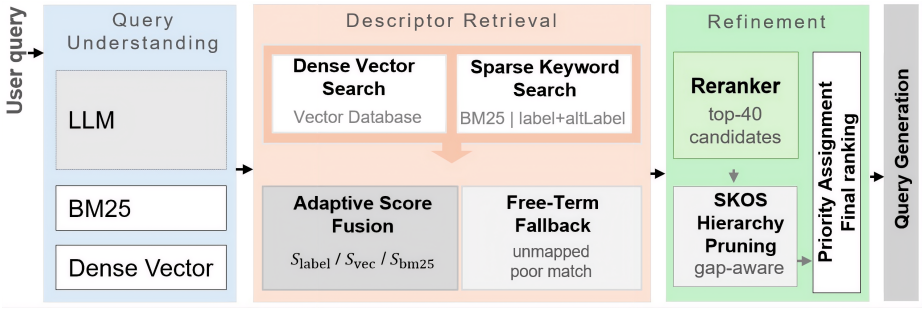

Fig. 2. System architecture of mapping free-text queries to thesaurus descriptors.

STW descriptor mapping, including topical concepts, geographic entities, and person names. We evaluated several locally deployable open-weight models for this task, including Mistral-NeMo 12B, GPT-OSS 20B, Qwen3.5 (2B, 9B, and 35B), Qwen3.6 35B, and LLaMA 3.2 3B. On the title-as-query benchmark using STW-assigned descriptors as ground truth (Section 5.1), all models achieved broadly comparable extraction performance, reflecting the constrained nature of the task. Larger models, particularly Qwen3.5 and Qwen3.6 35B, showed modest gains on noisier or less direct formulations, while models above 35B parameters were excluded due to hardware constraints. Qwen3.6 35B is therefore adopted as the primary extraction model.

To improve robustness of the extracted representations, prompt constraints and rule-based post-processing are applied to mitigate over-generation of generic academic terms (e.g., *impact*, *analysis*), unless explicitly relevant to query intent. Additional safeguards recover missing geographic entities, remove duplicated person-name fragments, and suppress topical extraction in author-centric queries to avoid introducing irrelevant retrieval constraints.

It is worth noting that even lighter models, including LLaMA 3.2 3B, provide a viable alternative under stricter latency or memory constraints, with only a modest performance degradation compared to Qwen3.6 35B (P@k 0.621 vs. 0.644, R@k 0.419 vs. 0.464, Macro-F1@k 0.458 vs. 0.503). Concept extraction uses temperature = 0.1, 150-token budget and JSON-decoding; the reranker [50] scores query–descriptor pairs via log-probability of a single next token.

Temporal constraints are extracted independently using a rule-based parser covering relative expressions (e.g., *recent*, *last five years*), absolute ranges, and directional constraints (e.g., *after 2020*). These are translated directly into Solr date filters, ensuring temporal scope is enforced at the index level. A domain-specific abbreviation map covering common economics and policy terminology (e.g., GDP, FDI, AI) is applied before and after LLM extraction to improve concept consistency and alignment with STW terminology.

Three-path concept extraction. Query understanding operates through three complementary extraction paths that run in parallel, each targeting a distinct failure mode of single-path retrieval.

LLM-decomposed path. The LLM extracts an explicit list of topical concepts from the query. Each concept is subsequently embedded and matched against the STW descriptor index via dense vector search. This path handles paraphrased and semantically indirect queries well, but may omit or rephrase terms whose exact label appears in the query verbatim.

Direct BM25 path. The raw user query is searched directly against the BM25 index built over STW preferred and alternative labels, bypassing LLM concept decomposition. This recovers descriptors whose labels align with the original query string but were missed during concept extraction, for instance due to paraphrasing, abstraction, or omission by the LLM.

Direct dense-vector path. The raw query embedding is searched against the Qdrant descriptor vector index, capturing global semantic associations that may be missed when the query is decomposed into individual concepts.

All three paths contribute to a shared candidate pool. Descriptors retrieved by both the concept-decomposed path and at least one direct path are treated as *cross-path hits*: their score is set to the maximum of the contributing path scores and boosted by a small multiplicative factor ($\times 1.05$, capped at 1.0), reflecting convergent evidence. The reranker remains the primary quality filter and may discard irrelevant cross-path hits. Remaining direct-path candidates are appended and tagged with their retrieval origin (`direct_bm25`, `direct_vector`, or `direct_hybrid`) for post-hoc analysis.

4.2 Phase 2: Descriptor Retrieval (mapping)

Per-concept lexical retrieval. For each concept extracted in Phase 1, BM25 retrieval [37] is performed over STW preferred labels, alternative labels, and scope notes. This complements the direct-query BM25 path from Phase 1, which operates on the full user query, whereas per-concept retrieval targets atomic semantic units isolated by the LLM.

Dense semantic retrieval. Each extracted concept is embedded using a multilingual dense retrieval model and matched against a precomputed STW descriptor vector index via approximate nearest-neighbour search. Retrieval is executed independently per concept, enabling parallelisation across the full concept set.

Candidate consolidation. Candidates from per-concept lexical retrieval, per-concept dense retrieval, and direct-query retrieval (Phase 1) are merged into a unified candidate pool via the cross-path fusion strategy described in Section 4.1 and passed to adaptive score fusion.

Adaptive score fusion. Each candidate descriptor is assigned a combined relevance score:

$$\hat{S} = w_{\text{vec}} \cdot S_{\text{vec}} + w_{\text{label}} \cdot S_{\text{label}} + w_{\text{bm25}} \cdot S_{\text{bm25}}$$

where S_{vec}, S_{label}, and S_{bm25} denote dense similarity, label overlap, and BM25 score respectively. The weight triple ($w_{\text{vec}}/w_{\text{label}}/w_{\text{bm25}}$) is selected from four schemes in priority order, ensuring that lexical evidence takes precedence when both label and vector signals are strong:

- *keyword-dominant* (0.15 / 0.75 / 0.10): $S_{\text{label}} \geq 0.95$;
- *keyword-assisted* (0.40 / 0.45 / 0.15): $S_{\text{label}} \geq 0.75$;
- *vector-dominant* (0.75 / 0.15 / 0.10): $S_{\text{vec}} \geq 0.80$;
- *balanced* (0.55 / 0.25 / 0.20): otherwise.

The label thresholds distinguish near-exact (≥ 0.95) from high-confidence partial matches (≥ 0.75); the vector threshold (≥ 0.80) corresponds to approximately the 90th percentile of cosine similarities on a calibration subset, above which candidates are near-synonymous with the query concept. Informal checks on the reranker margin and fallback threshold suggest that macro-F1 varies by less than 0.01 under ± 0.05 perturbations for the configurations tested. A more exhaustive sensitivity analysis across all threshold parameters remains future work.

Free-term fallback. When no STW descriptor sufficiently covers an extracted concept, or when the best-matching descriptor covers less than 40% of the concept's content words, the original concept is retained as a free-text retrieval term. This preserves recall for emerging or non-standard terminology not yet represented in STW.

4.3 Phase 3: Descriptor Refinement

Neural reranking. The top-40 candidate descriptors are re-scored using Qwen3-Reranker-4B [50], deployed locally via Ollama. Each query–descriptor pair is framed as a primary-subject classification task, where the model estimates whether the descriptor represents a central topical focus rather than incidental context.

Scores are derived by aggregating first-token log-probabilities over affirmative and negative response variants and applying a yes/no softmax, yielding a calibrated relevance probability $p_{\text{yes}} \in [0, 1]$. Empirically, we found that positive primary-subject framing produced substantially better score separation than negatively framed alternatives, enabling reliable threshold-based filtering (see Section 5.1 for calibration details).

To prevent over-filtering, the system enforces a minimum retention of two descriptors and applies an adaptive threshold $\max(\tau_{\text{base}},\ s_{\max} - 0.35)$, where $s_{\max}$ denotes the highest reranker score in the candidate set. This ensures retention of descriptors that remain sufficiently close to the top-ranked candidate.

SKOS-aware hierarchy pruning. A symbolic refinement step removes redundant broader descriptors when narrower descendants are already selected. Hierarchical relations are resolved directly from descriptor metadata. This addresses a retrieval-specific failure mode of Boolean controlled-vocabulary search, where simultaneous activation of semantically nested descriptors can unnecessarily reduce recall in conjunctive Boolean retrieval.

Priority assignment. The final descriptor set is partitioned into priority and non-priority descriptors. Priority descriptors form the initial Boolean query, while lower-ranked descriptors remain available for optional activation during interaction. A bounded number of priority descriptors is enforced to maintain tractable result-set sizes.

4.4 Phase 4: Interactive Query Construction and User Interface

This phase operationalises the retrieval model into an interactive user interface, as shown in Figure 3.

Interactive retrieval control. Rather than treating query formulation as a one-shot transformation, the system exposes STW descriptor mappings as interactive retrieval controls. Users can inspect, activate, deactivate, or expand descriptors while preserving compatibility with the underlying Boolean retrieval model.

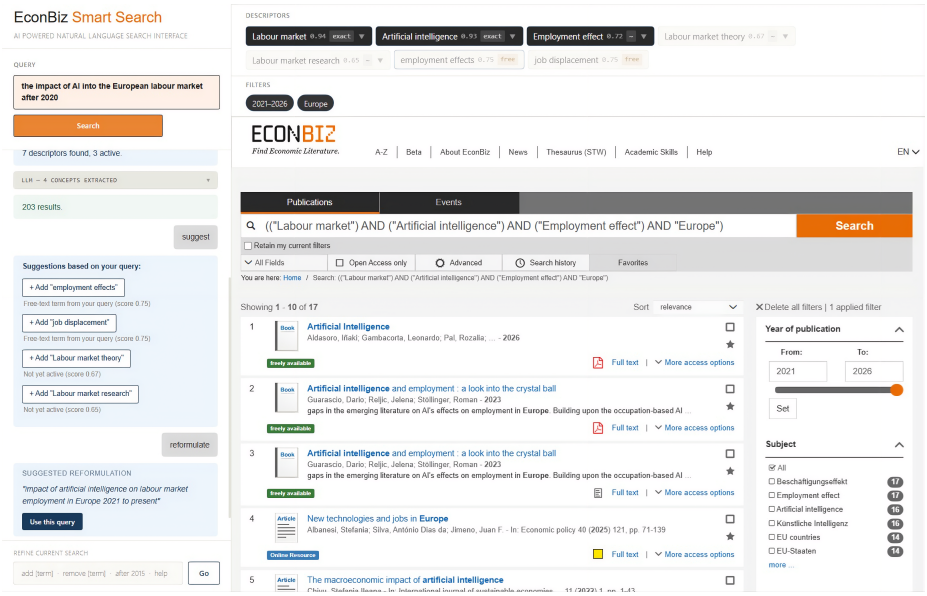

Fig. 3. Prototype Smart Search interface for EconBiz, exposing STW descriptors and a chat-based control layer for iterative query refinement.

Boolean query generation. Priority descriptors are assembled into a Boolean Solr query in which each descriptor contributes an OR-group over its preferred label and user-activated alternatives, while descriptor groups are combined conjunctively. Geographic and temporal constraints are appended as structured filters, preserving compatibility with the existing EconBiz retrieval infrastructure.

Progressive interface and descriptor interaction. Lexical BM25 results are returned immediately after query submission, providing early feedback, while semantic descriptor processing, reranking, and pruning proceed asynchronously. Descriptor modifications update the retrieval state without re-executing the full LLM pipeline. Expanding a descriptor retrieves semantically related alternatives from STW labels and dense-neighbour retrieval, supporting controlled query expansion within the thesaurus structure.

Conversational control and failure recovery. A constrained chat interface exposes a set of operations, such as: *reformulate*, which rewrites queries into STW-aligned forms, and *suggest*, which proposes descriptor additions or substitutions based on confidence scores and result counts. This enables incremental relaxation of over-constrained Boolean queries and recovery from empty result sets, while maintaining grounding in controlled vocabulary structure.

5 Evaluation

The evaluation examines two complementary aspects of the proposed system: (i) descriptor mapping quality under controlled offline conditions, and (ii) recovery of real-world zero-result retrieval failures via interactive descriptor refinement.

5.1 Title-as-Query Evaluation

Descriptor mapping quality is evaluated using a title-as-query proxy, where publication titles are treated as input queries and cataloguer-assigned STW descriptors serve as ground truth. This evaluation strategy has precedent in controlled-vocabulary retrieval research because it enables exact descriptor-level comparison at scale while preserving concise natural-language input representations [40]. Although publication titles are more formal than real user queries, they provide a controlled approximation of vocabulary mismatch effects observed in Section 3.4.

The evaluation dataset comprises 1,000 EconBiz records sampled across major STW top-level categories. Records with fewer than two or more than ten assigned descriptors were excluded to avoid under-specified or excessively broad indexing cases. We report macro-averaged Precision@k, Recall@k, and F1@k, where k varies per query; Avg. k in Table 1 is its mean across queries.

In addition to benchmarking, this evaluation was used for system calibration. Reranker thresholds, descriptor selection margins, hierarchy-pruning parameters, and priority activation rules were tuned on a held-out subset and refined through qualitative inspection of real-user failure cases.

To isolate the contribution of individual pipeline components, retrieval performance was evaluated under both LLM-only concept extraction and the proposed three-path fusion extraction strategy:

- **BM25 only**: lexical retrieval over STW labels and scope notes;
- **Dense only (BGE-M3)**: vector retrieval using BGE-M3 [8];
- **Dense only (Qwen3-Embedding)**: vector retrieval using Qwen3-Emb. [50];
- **Hybrid**: lexical–semantic fusion without reranking or pruning;
- **Full pipeline**: hybrid retrieval with reranking, SKOS-aware pruning, and priority assignment.

Table 1 shows that dense retrieval substantially improves recall over lexical BM25 retrieval, particularly when using Qwen3-Embedding. However, BM25 remains comparatively precise for exact descriptor matches, confirming the complementary strengths of semantic and lexical retrieval signals.

Hybrid retrieval substantially improves overall descriptor mapping quality, confirming that lexical and dense retrieval capture partially distinct relevance signals. The full pipeline further improves precision and macro-F1 under both extraction settings, reducing the mean descriptor count from 4.3 to 2.6 under three-path fusion — operationally important for Boolean retrieval where excessive descriptor activation over-constrains result sets. The limited recall decrease suggests that reranking and SKOS-aware pruning remove low-value or redundant descriptors without harming topical coverage.

Table 1. Descriptor mapping performance on the title-as-query benchmark (N = 1,000). Avg.k denotes the mean number of priority descriptors per query.

Condition	P@k	R@k	Macro-F1@k	Avg. k
BM25 only	0.244	0.348	0.270	4.9
Dense only (BGE-M3)	0.200	0.303	0.231	5.0
Dense only (Qwen3)	0.278	0.436	0.327	5.0
LLM concept extraction only				
Hybrid retrieval (without reranker)	0.512	0.379	0.419	2.6
Full pipeline (LLM-only)	**0.644**	0.367	0.463	2.1
Three-path fusion extraction				
Hybrid retrieval (without reranker)	0.432	**0.510**	0.447	4.3
Full pipeline (fusion)	0.619	0.464	**0.503**	2.6

Three-path fusion extraction shifts the precision–recall balance relative to LLM-only extraction: recall increases from 0.367 to 0.464 while precision remains strong at 0.619, yielding the highest macro-F1 (0.503) across all evaluated configurations. This improvement reflects the broader candidate pool introduced by direct-query retrieval paths, which recover descriptors missed during LLM concept decomposition, at the cost of admitting additional candidates.

The evaluation does not directly measure document-level retrieval effectiveness or end-user satisfaction. The title-as-query proxy nonetheless provides a scalable, controlled benchmark for comparing descriptor mapping strategies, complemented by the manual evaluation in Section 5.2 and discussed further in Section 6.

5.2 Manual Evaluation on Zero-Result Queries

To complement the offline benchmark with real-world retrieval failures, we manually evaluated 50 zero-result queries sampled from EconBiz interaction logs. Queries were drawn from the long-query segment (>8 words), identified in Section 3.4 as the highest-failure category. Two independent assessors evaluated each query across six dimensions: Descriptor Correctness, Priority Descriptor Quality, Filter Correctness, Success@5, Recovery T1 (after one descriptor toggle), and the number of interactions required for recovery. With backgrounds in economics, social sciences, or library science and uninvolved in system development, they rated descriptor appropriateness by weighing specificity, synonymy, and hierarchical fit against the query and STW vocabulary. This process necessarily involves interpretive judgement, particularly for specialised terminology and semantically adjacent STW concepts, so the evaluation should be understood as a pragmatic assessment of retrieval usefulness rather than authoritative validation. Disagreements were resolved through discussion to produce a unified set.

Descriptor Correctness received correct or partially correct ratings for 44 of 50 queries (88.0%), indicating that the pipeline reliably identifies relevant STW concepts for the large majority of complex zero-result queries. Priority Descriptor Quality achieved correct or partial ratings for 42 queries (84.0%), suggesting

that descriptor ranking and prioritisation are generally consistent with concept-level correctness. To assess consistency between concept-level and priority-level descriptor assessments, we computed weighted Cohen's κ across all 50 queries, yielding $\kappa = 0.68$, indicating substantial agreement. Filter correctness was assessed on 21 queries with explicit geographic or temporal constraints, of which 71.4% were assigned correct or partially correct filters. Initial Success@5 without interaction was achieved for 30 queries (60.0%). A further 11 queries (22.0%) achieved Success@5 after a single descriptor toggle (Recovery T1), yielding successful retrieval for 82.0% of queries with at most one interaction. Queries requiring additional interaction had a mean of 1.7 toggles to recovery.

These results support the system's central operational claim: zero-result queries are recoverable via descriptor-level refinement, particularly for diagnosing and relaxing over-constrained Boolean retrieval formulations.

6 Discussion

6.1 Design Decisions and Operational Trade-offs

Several architectural choices were shaped by operational constraints specific to digital-library deployment. First, concept extraction relies on locally deployed language models rather than proprietary external systems to satisfy institutional privacy requirements and ensure fully local query processing. Although larger proprietary models may achieve slightly higher extraction quality, the evaluation shows that smaller local models provide sufficient semantic fidelity when combined with structured prompting and rule-based post-processing. Second, reranking was formulated as a positive primary-subject classification task based on empirical calibration results. Negatively framed prompts produced compressed score distributions with limited discriminative power, whereas affirmative formulations yielded clearer score separation and more stable threshold behaviour. This indicates that prompt formulation strongly influences the calibration behaviour of instruction-tuned rerankers used as lightweight relevance classifiers. Third, SKOS-aware hierarchy pruning addresses a retrieval constraint specific to Boolean descriptor search: simultaneously activating semantically nested descriptors can unnecessarily reduce recall in conjunctive retrieval settings. Exploiting explicit thesaurus structure therefore improves retrieval behaviour while preserving semantic specificity. Taken together, these findings highlight the value of hybrid retrieval architectures in digital libraries: neural components provide semantic flexibility, while controlled-vocabulary structure maintains consistency with established indexing practice.

6.2 Limitations and Future Work

Several limitations should be considered when interpreting the present results.

First, operational thresholds, including reranker confidence cutoffs and hierarchy-pruning margins, were calibrated empirically on the title-as-query benchmark. A

more exhaustive sensitivity analysis across threshold parameters remains future work, as does a per-rule ablation of the safeguards in Section 4.1. Qualitatively, candidate-pool generation in Phase 1 (including the generic-term filter) and the reranker confidence cutoffs were the most consequential stages for precision.

Second, the title-as-query benchmark evaluates descriptor mapping quality rather than downstream document-level retrieval effectiveness. While appropriate for controlled comparisons of mapping strategies, additional measures such as NDCG, task completion, or interactive success metrics would provide a more comprehensive evaluation. Third, manual evaluation of zero-result queries involves interpretive judgement. In economics, where conceptual boundaries are subtle, descriptor assignment requires domain expertise and contextual interpretation. Our assessments therefore reflect bounded plausibility judgements rather than authoritative cataloguing decisions. Independent assessors and agreement statistics mitigate but do not eliminate this subjectivity. Fourth, STW coverage is necessarily finite; free-term fallback preserves recall for emerging or specific topics but may reduce alignment with the controlled vocabulary. Finally, only 32% of EconBiz records contain at least one STW descriptor, limiting the direct impact of descriptor-guided retrieval across the full collection. Taken together, these factors indicate a component-level validation rather than definitive evidence of end-user retrieval improvement.

A controlled A/B evaluation against the standard EconBiz search interface is planned once system parameters are stabilised, assessing task success, query reformulation behaviour, and user satisfaction under realistic search conditions. To support large-scale evaluation, we will leverage the STELLA framework integrated into the EconBiz environment for experimental scenarios [6], enabling systematic analysis under operational conditions. Future work includes domain-adaptive embedding fine-tuning, improved multilingual extraction and reranking, and descriptor expansion leveraging associative STW relations.

7 Conclusion

This paper presented a deployable multi-stage neuro-symbolic pipeline for mapping natural-language user queries to STW descriptors in the EconBiz digital library, addressing a measurable retrieval failure caused by vocabulary mismatch between natural-language queries and controlled-vocabulary indexing. Combining local LLM-based concept extraction, hybrid lexical–semantic retrieval, adaptive score fusion, neural reranking, and SKOS-aware hierarchy pruning, the approach substantially improves descriptor mapping quality over lexical-only and dense-only baselines while remaining fully compatible with existing Solr-based infrastructure.

More broadly, the results suggest that hybrid architectures combining neural semantic modelling with explicit thesaurus structure offer a practical, infrastructure-preserving path toward interactive, explainable retrieval in controlled-vocabulary digital libraries.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of AMIA Annual Symposium, pp. 17–21 (2001)
2. Auer, S., et al.: DBpedia: A nucleus for a web of open data. In: Proc. ISWC 2007, LNCS, vol. 4825, pp. 722–735. Springer (2007) https://doi.org/10.1007/978-3-540-76298-0_52
3. Azad, H.K., Deepak, A.: Query expansion techniques for information retrieval: A survey. Information Processing and Management 56(5), 1698–1735 (2019) https://doi.org/10.1016/j.ipm.2019.05.009
4. Belkin, N.J., Cool, C., Kelly, D., Lin, S.-J., Park, S.Y., Perez-Carballo, J., Sikora, C.: Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. Information Processing & Management 37(3), 403–434 (2001) https://doi.org/10.1016/S0306-4573(00)00055-8
5. Binding, C., Tudhope, D.: Improving interoperability using vocabulary linked data. Int. J. Digital Libraries 17(1), 5–21 (2016) https://doi.org/10.1007/s00799-015-0166-y
6. Breuer, T., Schaer, P., Tavakolpoursaleh, N., Schaible, J., Wolff, B., Müller, B.: STELLA: Towards a Framework for the Reproducibility of Online Search Experiments. Proceedings of the Open-Source IR Replicability Challenge co-located with the 42nd ACM SIGIR Conference on Research and Development in Information Retrieval. CEUR Workshop Proceedings, vol. 2409, pp. 8–11 (2019)
7. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computing Surveys 44(1), 1–50 (2012) https://doi.org/10.1145/2071389.2071390
8. Chen, J., et al.: BGE M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Findings of ACL 2024, pp. 2318–2335 (2024) https://doi.org/10.18653/v1/2024.findings-acl.137
9. Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H.: The ontology lookup service: A lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics 7, 97 (2006) https://doi.org/10.1186/1471-2105-7-97
10. Diaz, F., Mitra, B., Craswell, N.: Query expansion with locally-trained word embeddings. In: Proc. ACL 2016, pp. 367–377 (2016) https://doi.org/10.18653/v1/P16-1035
11. EconBiz. https://www.econbiz.de/ (accessed April 2026)
12. Esposito, M., et al.: Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Information Sciences 514, 88–105 (2020) https://doi.org/10.1016/j.ins.2019.12.002
13. Formal, T., Piwowarski, B., Clinchant, S.: SPLADE: Sparse lexical and expansion model for first stage ranking. In: Proc. 44th ACM SIGIR, pp. 2288–2292 (2021) https://doi.org/10.1145/3404835.3463098
14. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Commun. ACM 30(11), 964–971 (1987) https://doi.org/10.1145/32206.32212
15. Golub, K.: Automated subject indexing: An overview. Cataloging and Classification Quarterly 59(8), 702–719 (2021) https://doi.org/10.1080/01639374.2021.2012311

16. Gross, T., Taylor, A.G., Joudrey, D.N.: Still a lot to lose: the role of controlled vocabulary in keyword searching. Cataloging & Classification Quarterly 53(1), 1–39 (2015) https://doi.org/10.1080/01639374.2014.917447
17. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. Int. J. Human-Computer Studies 43(5–6), 907–928 (1995) https://doi.org/10.1006/ijhc.1995.1081
18. Hajra, A., Tochtermann, K.: Linking science: Approaches for linking scientific publications across different LOD repositories. International Journal of Metadata, Semantics and Ontologies 12(2–3), 124–141 (2017) https://doi.org/10.1504/IJMSO.2017.090778
19. Hajra, A.: Enhancing academic search with agentic RAG and knowledge graphs: A case study on EconBiz. In: Rocha, A., García-Peñalvo, F., Costa, C.J., Gonçalves, R. (eds.) Proceedings of the 20th Iberian Conference on Information Systems and Technologies (CISTI). LNNS, vol. 1717, pp. 907–918. Springer, Cham (2026) https://doi.org/10.1007/978-3-032-10721-3_77
20. Hitzler, P., et al.: Neuro-symbolic approaches in artificial intelligence. National Science Review 9(6), nwac035 (2022) https://doi.org/10.1093/nsr/nwac035
21. ISO 25964-1: Information and Documentation – Thesauri and Interoperability with Other Vocabularies – Part 1: Thesauri for Information Retrieval. ISO, Geneva (2011)
22. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7(3), 535–547 (2021) https://doi.org/10.1109/TBDATA.2019.2921572
23. Kasprzik, A.: Automating subject indexing at ZBW: Making research results stick in practice. LIBER Quarterly 33(1) (2023) https://doi.org/10.53377/lq.13579
24. Khattab, O., Zaharia, M.: ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In: Proc. 43rd ACM SIGIR, pp. 39–48 (2020) https://doi.org/10.1145/3397271.3401075
25. Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). 2020, pp. 9459–9474 (2020) https://doi.org/10.48550/arXiv.2005.11401
26. Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proc. 44th Int. ACM SIGIR Conf., pp. 2356–2362 (2021) https://doi.org/10.1145/3404835.3463238
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. ICLR (2013) https://doi.org/10.48550/arXiv.1301.3781
28. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. W3C Recommendation (2009)
29. MTEB Leaderboard. https://huggingface.co/spaces/mteb/leaderboard (accessed Jan 2026)
30. Neubert, J.: Bringing the "Thesaurus for Economics" on to the Web of Linked Data. LDOW Workshop (2009)
31. Nogueira, R., Cho, K.: Passage Re-ranking with BERT. arXiv:1901.04085 (2019) https://doi.org/10.48550/arXiv.1901.04085
32. Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document ranking with a pretrained sequence-to-sequence model. In: Findings of EMNLP 2020, pp. 708–718 (2020)
33. Noy, N.F., McGuinness, D.L.: Ontology Development 101. Stanford KSL Technical Report (2001)

34. Oelen, A., Jaradeh, M.Y., Auer, S.: ORKG ASK: A neuro-symbolic scholarly search and exploration system. arXiv preprint arXiv:2412.04977 (2024)
35. Ollama. https://ollama.com (accessed Jan 2026)
36. Qdrant Vector Database. https://qdrant.tech (accessed Jan 2026)
37. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval 3(4), 333–389 (2009) https://doi.org/10.1561/1500000019
38. Rogers, A., Kovaleva, O., Rumshisky, A.: A Primer in BERTology: What We Know About How BERT Works. Trans. Assoc. Comput. Linguistics 8, 842–866 (2020) https://doi.org/10.1162/tacl_a_00349
39. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975) https://doi.org/10.1145/361219.361220
40. Savoy, J.: Bibliographic database access using free-text and controlled vocabulary. Inf. Process. Manag. 41(4), 873–890 (2005) https://doi.org/10.1016/j.ipm.2004.01.004
41. Sołoducho-Pelc, L., Sulich, A.: Artificial intelligence in academic research: A comparative study of Scopus AI and Web of Science Research Assistant. Procedia Computer Science 270, 6126–6135 (2025) https://doi.org/10.1016/j.procs.2025.10.082
42. Srinivasan, P.: Query expansion and MEDLINE. Information Processing & Management 32(4), 431–443 (1996) https://doi.org/10.1016/0306-4573(95)00076-3
43. STW Thesaurus for Economics. https://zbw.eu/stw/version/latest/about.en.html (accessed April 2026)
44. Taylor, J., Dagan, K., Youngberg, M., Kaufman, T., Radding, J.: A Survey of AI tools in Library Tech: Accelerating into and Unlocking Streamlined Enhanced Convenient Empowering Game-Changers. Journal of Electronic Resources Librarianship, 37(2), 217–229. (2025) https://doi.org/10.1080/1941126X.2025.2497738
45. Thakur, N., et al.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. arXiv:2104.08663 (2021) https://doi.org/10.48550/arXiv.2104.08663
46. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proc. 17th Annual ACM SIGIR Conf., pp. 61–69 (1994) https://doi.org/10.1007/978-1-4471-2099-5_7
47. Yang, X.W., et al.: Neuro-symbolic artificial intelligence. arXiv:2508.13678 (2025) https://doi.org/10.48550/arXiv.2508.13678
48. ZBW. https://www.zbw.eu/en (accessed April 2026)
49. Zerhoudi, S., Granitzer, M.: Comparative Analysis: User Interactions in Public and Private Digital Libraries Datasets. In: TPDL 2024, LNCS vol. 15178, pp. 162–172 (2024) https://doi.org/10.1007/978-3-031-72440-4_16
50. Zhang, Y., Li, M., Long, D., et al.: Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models arXiv:2506.05176 (2025) https://doi.org/10.48550/arXiv.2506.05176
51. Zhao, W.X., Zhou, K., Li, J., et al.: A survey of large language models. Frontiers of Computer Science 20, 2012627 (2026) https://doi.org/10.1007/s11704-026-60308-3