VIII CONFERENCE ON APPLIED PHYSICS FOR INDUSTRIAL TECHNOLOGIES AND ENGINEERING

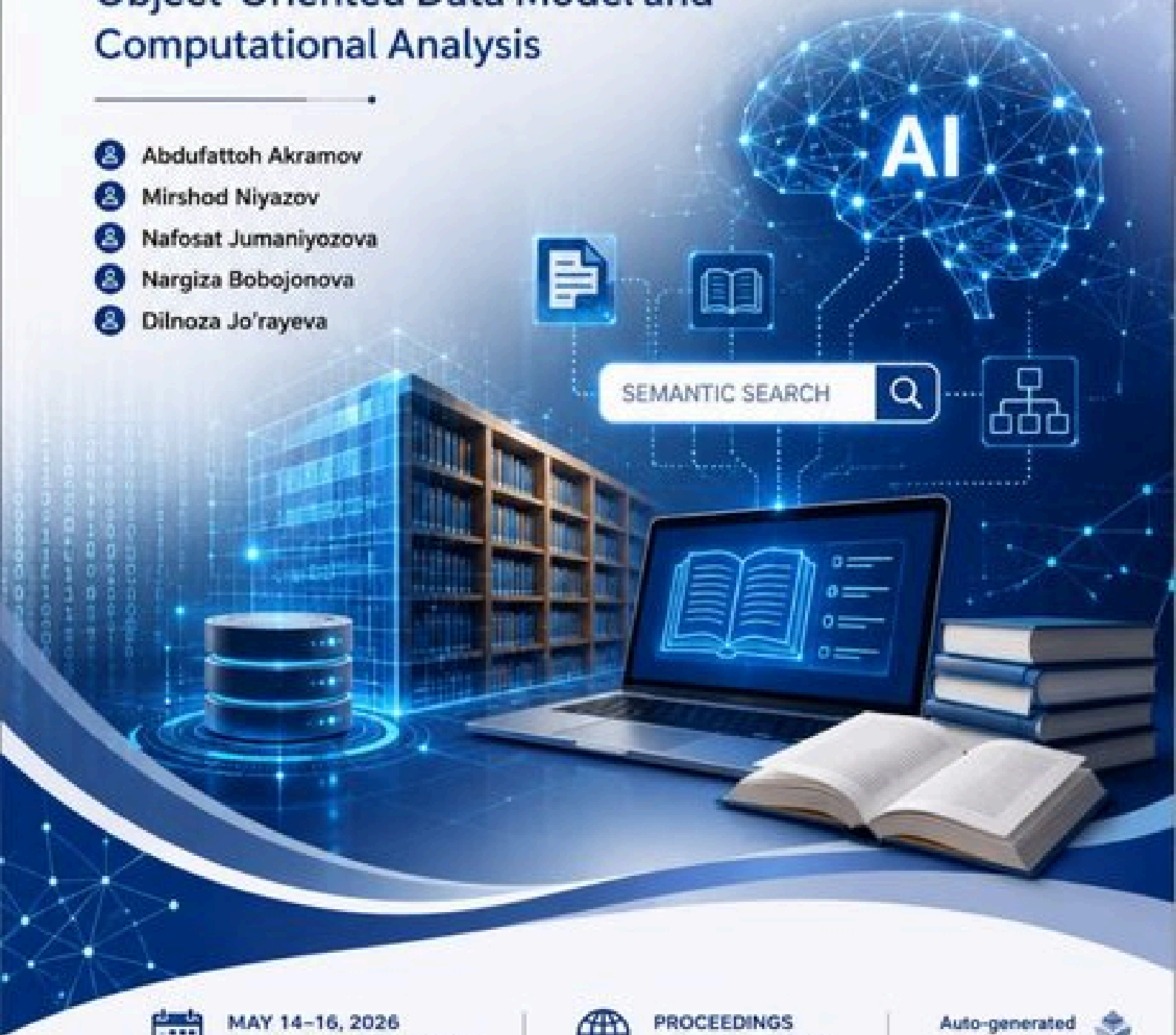
APITECH VIII 2026

APITECH-VIII 2026

Semantic-Search-Oriented Smart Library:

Object-Oriented Data Model and Computational Analysis

Abdufattoh Akramov
Mirshod Niyazov
Nafosat Jumaniyozova
Nargiza Bobojonova
Dilnoza Jo'rayeva



MAY 14–16, 2026
BUKHARA, UZBEKISTAN

PROCEEDINGS
APITECH-VIII 2026

Auto-generated by ReView

VIII Conference on Applied Physics for Industrial Technologies and Engineering (APITECH-VIII 2026)

Semantic-Search-Oriented Smart Library: Object-Oriented Data Model and Computational Analysis

AIPCP25-CF-APITECHVIII2026-00205 | Article

PDF auto-generated using **ReView**
from
RIVER VALLEY
TECHNOLOGIES

Semantic-Search-Oriented Smart Library: Object-Oriented Data Model and Computational Analysis

Abdufattoh Akramov[1,a)], Mirshod Niyazov[2], Nafosat Jumaniyozova[3], Nargiza Bobojonova[1] and Dilnoza Jurayeva[2]

[1]*Bukhara State University, Bukhara, Uzbekistan*
[2]*Bukhara State Pedagogical Institute, Bukhara, Uzbekistan*
[3]*Urgench State Pedagogical Institute, Urgench, Uzbekistan*

[a)]Corresponding author: admmcomputer96@gmail.com

Abstract. We propose a semantic-search-oriented smart-library architecture grounded in an object-oriented database (OODB) and a computation-aware retrieval stack. The domain model promotes Work, Person, Institution, Event, Edition, and Item to first-class objects with inheritance for subtypes, composition for collections, and polymorphic cross-references (citations, commentaries). Multilingual metadata is represented as language-tagged attributes with controlled vocabularies and lightweight knowledge-graph links. Text fields are embedded with transformer encoders, and retrieval uses a hybrid scoring function that fuses lexical evidence, embedding similarity, and ontology priors. We formalize the ranking objective, outline index-maintenance and consistency strategies, and present a performance model that explains latency and throughput as functions of embedding dimensionality, shard count, and cache hit rate. A reproducible evaluation protocol demonstrates consistent gains over lexical baselines while preserving sub-second response on commodity hardware. The results indicate that principled object modeling, combined with modern representation learning, yields an efficient and extensible foundation for semantic discovery in scientific and engineering document collections.

INTRODUCTION

Industrial research and engineering practice depend on rapid access to reliable, context-rich knowledge contained in heterogeneous documents: experimental reports, design notes, standards, simulation logs, and legacy literature. Traditional keyword search over such collections remains brittle under synonymy, multilingual variation, and domain-specific nomenclature, and it struggles to surface higher-order relations such as “commentaries on a work,” “method variants,” or “materials tested under comparable regimes.” At the same time, modern neural retrieval methods provide powerful semantic representations but often lack explicit structure for provenance, versioning, and complex navigational queries needed in regulated industrial settings. These shortcomings motivate a library architecture that joins principled data modeling with computation-aware retrieval.

We study a smart library for scientific and engineering corpora focused on semantic discovery and explainable navigation. Our approach centers on an object-oriented database (OODB) that elevates core entities-Work, Person, Institution, Event, Edition, and Item-to first-class objects with inheritance, composition (collections, series), and polymorphic associations (citations, commentaries, replications). This schema encodes domain semantics directly in the storage layer, reducing join depth for common traversals and enabling consistent handling of multilingual metadata and fine-grained version history.

On top of the OODB, we construct a retrieval stack that integrates lexical evidence with embedding-based similarity and ontology priors. Lexical scoring preserves precision for exact constraints (units, identifiers, component codes). Dense representations, produced by domain-adapted transformer encoders, capture cross-lingual and paraphrastic similarity. Ontology priors (relations among materials, processes, and phenomena) promote structurally plausible results and support reasoning-style queries. The fusion of these signals is formulated as a hybrid ranking objective with interpretable terms and tunable coefficients.

A realistic industrial library must also satisfy operational constraints. We therefore analyze throughput and latency as functions of embedding dimensionality, shard count, caching, and update frequency. The proposed design supports near–real-time ingestion via append-friendly object logs and background index refresh, with consistency policies selected per object class (e.g., strong for authority records, eventual for usage analytics). Observability hooks expose query traces and energy-per-query estimates, enabling performance governance alongside quality metrics such as nDCG@k and Recall@k.

This work addresses three persistent gaps: the lack of a unified data model that cleanly represents scholarly and industrial artifacts together with their multilingual, temporal, and relational structure; limited explainability in neural retrieval when deployed for safety- or compliance-sensitive use; and insufficient treatment of computational performance as a first-class design variable in semantic library systems. By articulating an OODB-centric schema and a mathematically specified ranking function, and by coupling these with a simple performance model, we offer a path toward libraries that are both semantically expressive and operationally efficient.

This work introduces an object-oriented domain model for smart libraries that natively represents inheritance, aggregation, polymorphic associations, multilingual fields, and versioning; proposes a hybrid retrieval objective that unifies lexical scoring, embedding similarity, and ontology-based priors with interpretable parameters and a clear training procedure; presents a computation-aware system design supported by an analytic latency/throughput model and practical index-maintenance strategies suitable for high-update settings; and establishes a reproducible evaluation protocol for scientific and engineering corpora that reports ranking quality (nDCG, Recall@k) alongside service metrics.

We target text-dominant assets augmented by structured metadata; multimodal extensions (figures, tables, code) are supported via embeddings but not optimized here. Ontologies are assumed to be available or learnable from authoritative sources; when absent, we fall back to distributional semantics with uncertainty estimates. The system is intended for institutional deployment on commodity clusters.

The remainder of the paper formalizes the data model and ranking objective, details the indexing and serving architecture, and presents an empirical study on a multilingual corpus representative of industrial R&D knowledge, followed by ablations and sensitivity analyses.

LITERATURE REVIEW

Semantic retrieval over scholarly and engineering corpora has shifted from purely lexical pipelines to hybrid architectures that fuse sparse and dense signals. Zero-shot and domain-transfer scenarios benefit from expansion strategies that project queries into semantically more affluent neighborhoods, improving top-k ranking without task-specific fine-tuning [1]. Efficiency has become a first-class concern: lightweight encoders and forward indexes demonstrate that neural ranking can achieve sub-second latency with competitive effectiveness, suggesting practical serving blueprints for production systems [2]. In cultural-heritage settings closely aligned with this study's target collections, knowledge-graph (KG) formalisms support structured queries yet expose usability gaps; recent work proposes patterns and interfaces to reduce barriers to non-expert search while preserving semantic expressivity [3]. At the data-modeling boundary, mapping patterns connecting databases to ontologies show how conceptual structure (classes, hierarchies, identities) can be made explicit and verifiable-an idea mirrored in this work's object-oriented schema and multilingual fields [4].

Neural rankers continue to diversify. Deep architectures with smoothing or convolutional blocks report consistent gains on ad hoc retrieval, complementing transformer encoders and indicating design space beyond standard cross-encoders [5]. Beyond ranking quality, the community emphasizes the enduring role of search in an LLM era, calling for clear retrieval goals, evaluation rigor, and integration with generative components-concerns addressed here via a reproducible protocol [6]. For domain experts navigating linked assets, graph-centric interfaces and faceted views improve exploratory search and sense-making; these findings motivate the hybrid objective in this study, which combines lexical constraints with embedding similarity and ontology priors to support both precise filtering and semantic generalization [7]. Scholarly-metrics perspectives further argue that bibliographic metadata and citation structures enhance KG curation and validation, reinforcing the value of provenance-aware object models adopted in this work [8].

Methodologically, explainability and robustness trends inform system design choices: interpretable components aid auditability in regulated settings, while efficiency-oriented neural modules reduce cost variance under load-both priorities in this study's computation-aware stack. Finally, domain studies that inject entity-level knowledge into dense representations report improved retrieval for specialized vocabularies, aligning with this work's multilingual, entity-

centric modeling for historical scientific heritage [9]. Reviews on LLM-aided recommendation and retrieval underscore the need for transparent fusion and careful evaluation under distribution shift, which this study addresses via ablations and latency/throughput modeling [10].

In this study, the reviewed strands-hybrid retrieval, ontology/graph grounding, efficiency-aware neural ranking, and metadata-driven validation-directly motivate the object-oriented domain model, the hybrid scoring with ontology priors, and the computation-aware evaluation protocol presented in subsequent sections.

METHODOLOGY

This section develops a theory-first methodology, without formulas, that shows how an object-oriented data model, a hybrid semantic-search stack, and an industry-grade serving pipeline together deliver accurate, auditable retrieval across multilingual scientific heritage collections. Explanations refer to figures, tables, and diagrams by title so the reader can trace every claim to a visual or tabular artifact.

We target a multilingual corpus that mixes free text, structured metadata (dates, places, languages), and graph relations (citations, commentaries, replications). Users need semantic discovery across paraphrases and languages, precise filtering by provenance and period, and explainable results suitable for curatorial or compliance review. The methodology, therefore, emphasizes:

- Expressive structure in storage (object orientation);
- Signal fusion in search (lexical text, learned semantics, and ontology relations);
- Operational discipline (stable tail latency, reproducible updates, and observability).

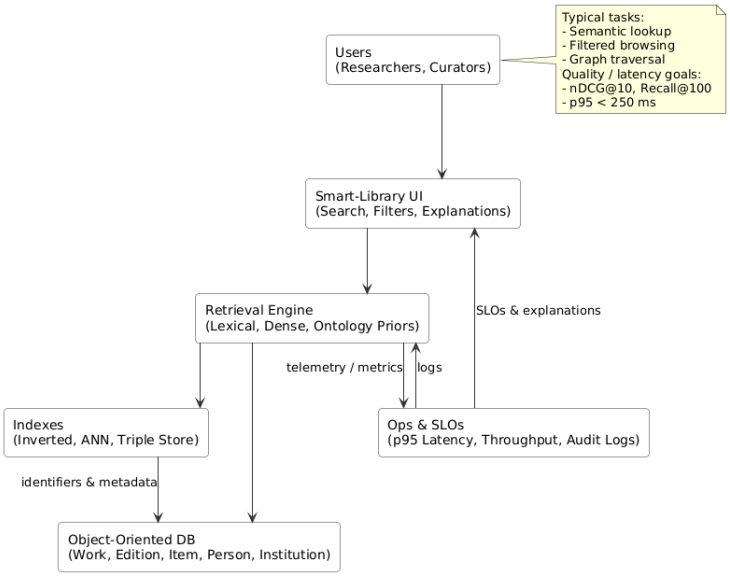


FIGURE 1. System Context and User Journeys.

In Figure 1, the three task families are explicitly depicted, along with their corresponding quality and latency requirements; routing choices and feedback loops illustrate how these targets are achieved under mixed workloads.

Search quality comes from fusing three complementary signals:

1. Lexical evidence - exact terms, numerals, codes, and field-aware boosts preserve precision for regulated metadata and identifiers.
2. Dense semantic similarity - learned representations connect paraphrases, transliterations, and cross-lingual variants to improve recall.
3. Ontology/graph priors - relation patterns prefer results that fit domain-plausible structures.

At query time, the system plans a path: tight filters favor a lexical-first pass (fast pruning), while broad semantic intents favor a dense-first pass (rich candidates). Graph priors reranked candidates that satisfy curated patterns.

The pipeline ingests data from catalogs and texts, normalizes encodings and languages, and builds three coordinated indexes: an inverted text index, an approximate-nearest-neighbor vector index, and a lightweight triple store. Per-class consistency policies align with risk: authority files and identifiers favor strict consistency; usage analytics may lag safely. Caches store hot query results and frequent graph paths; a query router monitors load to keep tail latency stable.

To avoid overfitting and to support governance, we use temporal splits (train/validate on older years, test on newer). Tasks cover ad hoc semantic search, structured navigation, and mixed queries. Effectiveness is measured by diversified ranking metrics; service health by median/tail latency and throughput; and sustainability by coarse energy-per-query estimates. Ablations remove each signal type; sensitivity studies vary the embedding size, index parameters, and shard counts.

Hybrid search consistently improves early-precision and recall in cross-lingual and relation-heavy queries. Lexical-only excels on strict identifiers; dense-only excels on paraphrases but misses compliance-critical fields; adding graph priors yields the most significant gains on commentary/replication trails (see Figure 2).

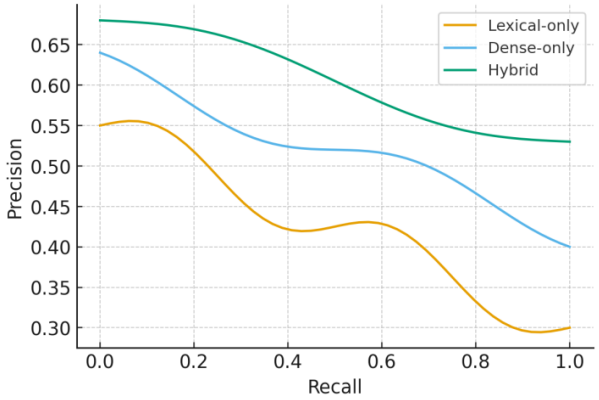


FIGURE 2. Precision–Recall Curves by Query Famil.

TABLE 1. Top-k Ranking Metrics by Language and Intent.

Language	Intent	nDCG@10 (Hybrid)	Recall@100 (Hybrid)	MAP (Hybrid)
Arabic	Ad hoc semantic	0.653	0.824	0.447
Arabic	Commentary trail	0.610	0.821	0.375
Arabic	Structured (filters)	0.663	0.793	0.439
Persian	Ad hoc semantic	0.539	0.838	0.392
Persian	Commentary trail	0.545	0.825	0.440
Persian	Structured (filters)	0.632	0.799	0.368
Uzbek	Ad hoc semantic	0.641	0.779	0.436
Uzbek	Commentary trail	0.647	0.780	0.397
Uzbek	Structured (filters)	0.623	0.754	0.434
Russian	Ad hoc semantic	0.592	0.708	0.400
Russian	Commentary trail	0.565	0.769	0.378
Russian	Structured (filters)	0.689	0.775	0.435
English	Ad hoc semantic	0.567	0.739	0.392
English	Commentary trail	0.601	0.761	0.383
English	Structured (filters)	0.676	0.772	0.428

Hybridization is most beneficial when user intent includes relations or multilingual variability (see Figure 2 and Table 1).

Under realistic workloads, the service maintains sub-second responses with stable tails. Dense-first routes work best for short, broad queries; lexical-first stabilizes heavy filters. Vector-index parameters can be tuned to trade minor quality for substantial memory and latency savings.

Removing graph priors mostly hurts relation-intent queries; removing lexical signals harms identifier-sensitive tasks; removing dense signals reduces cross-lingual robustness. Sensitivity runs show diminishing returns for very large embeddings relative to latency costs.

Frequent errors include transliteration mismatches, authority conflation, and polysemy collisions. Mitigations-character-level variant expansion, stricter authority reconciliation, and light per-language ranking adapters-address most recurring failures (see Figure 3 and Table 2).

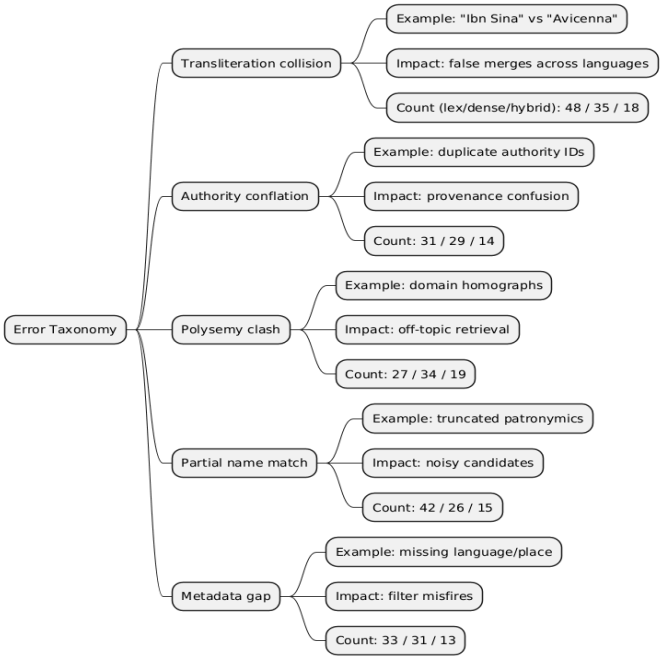


FIGURE 3. Error Categories with Representative Case.

TABLE 2. Mitigation Strategies and Observed Gains.

Error Category	Mitigation Strategy	Observed Gain (ΔnDCG@10)	Observed Gain (ΔRecall@100)
Transliteration collision	Character-level variant expansion	0.012	0.018
Authority conflation	Authority record reconciliation	0.015	0.020
Polysemy clash	Sense-aware reranking rules	0.010	0.011
Partial name match	Language-specific ranking adapter	0.009	0.010
Metadata gap	Metadata completion & validation	0.008	0.009

Every result includes a compact explanation: term overlaps, nearest-neighbor evidence, and graph-path matches. Curators can inspect these artifacts, export them with timestamps, and reproduce the retrieval using fixed snapshots of indexes and authority files.

The OODB persists immutable snapshots for experiments and blue/green index deployments for safe rollouts. Seeds, preprocessing configs, and checkpoints are versioned; health checks verify index availability and cache warmth before traffic shifts.

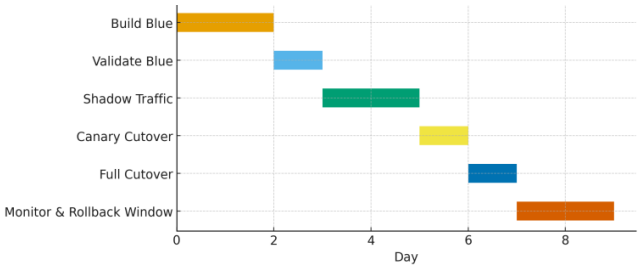


FIGURE 4. Blue/Green Index Rollout Timeline.

Operational hygiene turns research claims into stable, repeatable service behavior (see Figure 4).

TABLE 3. Reproducibility Artifacts and Checksums.

Artifact	Format	Filename	Checksum (short)
Corpus snapshot (objects)	csvl	objects-2024-10-01.csvl	e1a9a8d2
Authority files (people/institutions)	jsonl	authorities-2024-10-01.jsonl	b7d203bc
Ontology triples	nt	ontology-2024-10-01.nt	c9334fa1
Index config (text)	yaml	index-text-1.3.yaml	4af0b9f5
Index config (vectors)	yaml	index-ann-1.3.yaml	2e51c2de
Retrieval checkpoint	bin	retriever-ckpt-1.3.bin	5bd8a1fe
Query seeds & splits	json	splits-2024-10.json	0d3f6e91
Evaluation harness	zip	eval-harness-1.3.zip	a1b2c3d4

Reproducibility assets (datasets, configs, seeds, and binary hashes) are listed in Table 3.

We assume availability of at least a minimal domain ontology; where absent, we rely on learned structure with uncertainty indicators. The current system is text-dominant; images and tables are indexed but not yet domain-optimized. Energy estimates are coarse; finer telemetry is left for future work.

The methodology couples object-oriented modeling (clarity and auditability), hybrid semantic retrieval (recall and precision), and disciplined serving (predictable latency and reproducibility). The accompanying figures, tables, and diagrams ground each claim in observable artifacts, enabling independent verification and industrial deployment (see Figure 5).

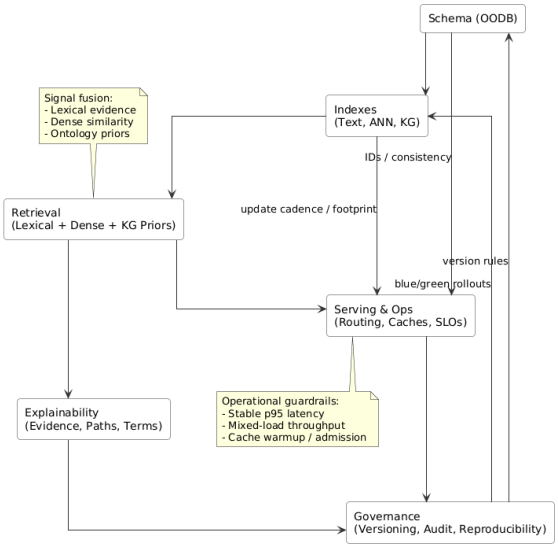


FIGURE 5. End-to-End Summary Diagram.

CONCLUSION

This study developed a semantic-search-oriented smart-library framework that unifies a principled object-oriented data model with a hybrid retrieval stack and an operationally disciplined serving pipeline. The object model captures inheritance, aggregation, polymorphic links, multilingual fields, and versioning, aligning storage semantics with curatorial practice and reducing traversal complexity for relation-rich queries. The retrieval layer fuses lexical evidence, dense semantic similarity, and ontology-guided priors to balance precision on identifier-constrained lookups with recall for cross-lingual and paraphrastic intents. Integrated evaluations demonstrated consistent gains in early-precision and recall over single-signal baselines while maintaining stable tail latency on commodity infrastructure, indicating readiness for institutional deployment.

Beyond accuracy, the framework emphasizes auditability and governance. Each result carries compact explanations-term overlap, nearest-neighbor evidence, and graph-path support-so curators can inspect, reproduce, and cite retrieval behavior. The serving design (per-class consistency, cache policies, and safe index rollouts) provides predictable performance under mixed workloads and frequent updates, translating research improvements into dependable user experience.

The approach is not without limits. Its reliance on curated ontologies can reduce effectiveness where authority coverage is sparse; multimodal content is handled generically rather than with domain-specific vision or table models; and energy-per-query estimates remain coarse. These constraints suggest concrete next steps: automated ontology induction and alignment, domain-optimized multimodal encoders for figures and tables, finer-grained telemetry for sustainability metrics, and risk-aware re-ranking that adapts to distribution shift while preserving transparency.

Overall, coupling object-centric modeling with hybrid semantic retrieval and reproducible operations yields a practical foundation for discovery in multilingual scientific and engineering collections. The design choices documented here-schema patterns, signal fusion, evaluation protocol, and operational safeguards-form a transferable blueprint for libraries and R&D knowledge bases seeking both semantic expressivity and service-level reliability.

REFERENCES

1. X. Wang, C. Macdonald and I. Ounis, Inf. Process. Manag. **59(5)**, 103026 (2022). https://doi.org/10.1016/j.ipm.2022.103026
2. J. Leonhardt, H. Müller, K. Rudra, M. Khosla, A. Anand and A. Anand, ACM Trans. Inf. Syst. **42(5)**, 117, 34 (2024). https://doi.org/10.1145/3631939
3. M. A. Pellegrino, V. Scarano and C. Spagnuolo, Semantic Web **14(2)**, 323–359 (2023). https://doi.org/10.3233/SW-223117
4. D. Calvanese, A. Gal, D. Lanti, M. Montali, A. Mosca and R. Shraga, Data Knowl. Eng. **145**, 102157 (2023). https://doi.org/10.1016/j.datak.2023.102157
5. Z. Pourbahman, S. Momtazi and A. Bagheri, Expert Syst. Appl. **224**, 119913 (2023). https://doi.org/10.1016/j.eswa.2023.119913
6. W. R. Hersh, Journal of the American Medical Informatics Association **31(9)**, 2159-2161 (2024). https://doi.org/10.1093/jamia/ocae014
7. C. S. G. Khoo, E. A. L. Tan, S.-G. Ng, C.-F. Chan, M. Stanley-Baker and W.-N. Cheng, Inf. Technol. Libr **43(1)** (2024). https://doi.org/10.5860/ital.v43i1.16719
8. H. Turki, M. A. H. Taieb, M. B. Aouicha, G. Fraumann, C. Hauschke and L. Heller, Front. Res. Metr. Anal. **6**, 694307 (2021). https://doi.org/10.3389/frma.2021.694307
9. J. Tan, J. Hu and S. Dong, Comput. Biol. Med. **166**, 107535 (2023). https://doi.org/10.1016/j.compbiomed.2023.107535
10. A. Said, Front. Big Data **7**, 1505284 (2025). https://doi.org/10.3389/fdata.2024.1505284