APPLIED DATA SCIENCE

ONTOLOGY ALIGNMENT CLASSIFIER

BASED ON NLP

Era Alcani
Francesco Dal Cero
Francesco Mastrosimone

Politecnico di Torino

Repertorio
UNLOCKING PATIENT PREFERENCES

TABLE OF CONTENTS

PROJECT OVERVIEW

01

PROJECT CONTEXT AND SOLUTION

THE CHALLENGE: SEMANTIC INCONSISTENCY

Context : Healthcare evidence is scattered across isolated "data silos".
Problem : Different terms (e.g., Fatigue vs. Asthenia) prevent meaningful integration of this data.

THE SOLUTION: AI SEMANTIC ENGINE

Goal : Develop an AI-powered engine to automatically harmonize fragmented data into structured evidence.
Impact : Replaces manual "brittle scripts" with a scalable, automated pipeline.

OBJECTIVES AND RESEARCH PERSPECTIVE

THE GOAL: AUTOMATED HARMONIZATION

To replace manual, brittle scripts with a robust framework that transforms raw attributes into high-confidence clinical evidence.

THE RESEARCH OBJECTIVES

SYSTEM LEVEL: Build a reusable framework for interpretable ontology alignment.

RESEARCH LEVEL: Evolve from simple string matching to capturing actual clinical meaning and context.

RESEARCH HYPOTHESES

02

HYPOTHESIS 1 — THE AMBIGUITY VS. EFFICIENCY

THE RESEARCH QUESTION

Do we always need expensive AI models like BERT to align clinical terms?

THE HYPOTHESIS

No. Complexity should match ambiguity. Deep learning is only necessary for ambiguous synonyms; deterministic rules are faster and equally accurate for the rest.

HYPOTHESIS 2 — ARCHITECTURE & SCALABILITY

THE RESEARCH QUESTION

How does the encoder architecture and search strategy impact alignment quality and scalability?

THE HYPOTHESIS

Efficiency demands a "Filter-then-Rank" architecture and scalability is only achieved by filtering candidates instantly before applying deep semantic scoring.

HYPOTHESIS 3 — ROBUSTNESS & DATASET CONSTRUCTION

THE RESEARCH QUESTION

How does dataset construction influence robustness and generalization in ontology alignment?

THE HYPOTHESIS

Training an AI on random data is insufficient for clinical accuracy. A model only generalizes well if it is explicitly taught to distinguish between tricky, similar-looking concepts.

HYPOTHESIS 4 — SEMANTIC UNDERSTANDING

THE RESEARCH QUESTION

Do Transformer-based models capture semantic equivalence better than lexical approaches?

THE HYPOTHESIS

Contextual Transformer-based models outperform lexical and embedding-only approaches in identifying semantic equivalence between concepts, especially when textual forms differ.

HYPOTHESIS 5 — BENCHMARK VALIDATION

THE RESEARCH QUESTION

How does the system perform against established ontology alignment benchmarks using standard evaluation frameworks?

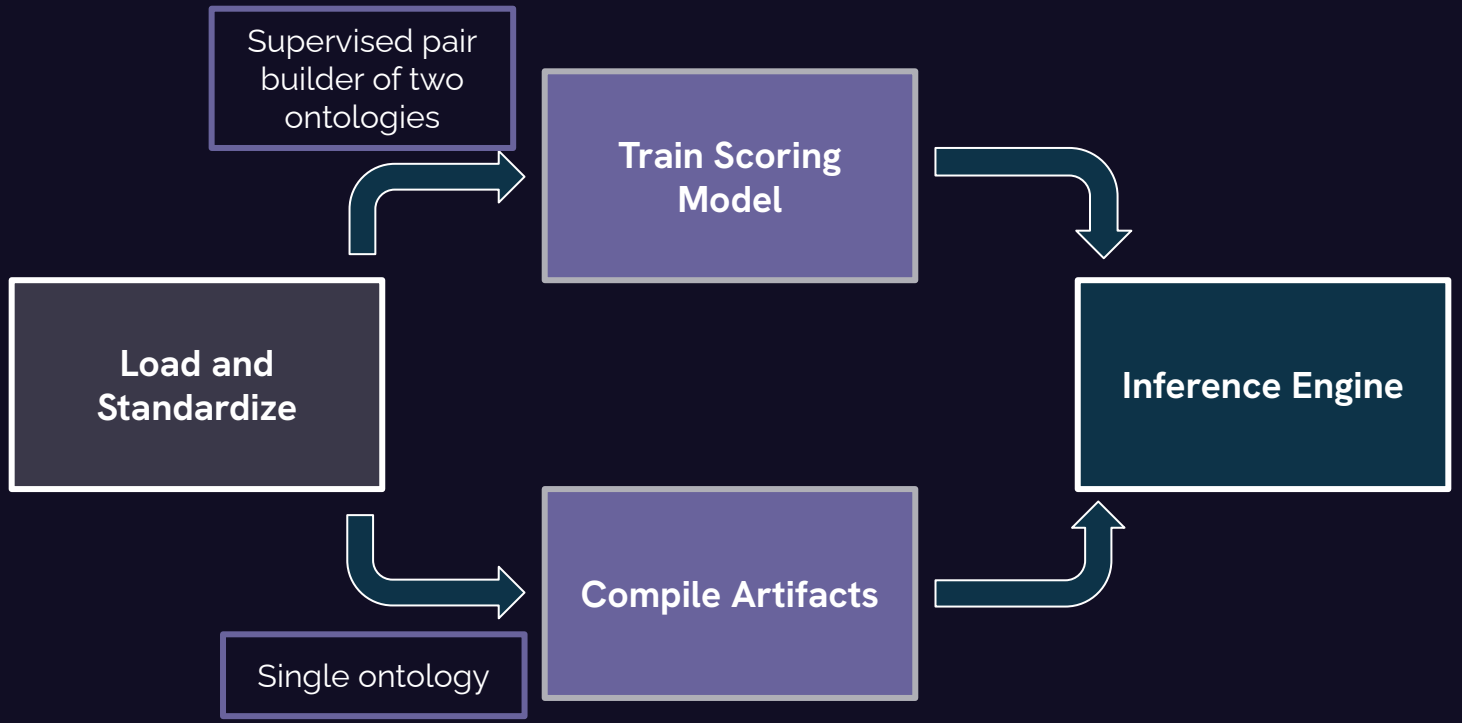
THE HYPOTHESIS

The proposed system achieves performance comparable to established ontology alignment benchmarks when evaluated using standard evaluation frameworks.

SYSTEM DESIGN

03

SYSTEM DESIGN - MODULAR OVERVIEW



- Reusable across studies – build once, run many times

SYSTEM WORKFLOWS & ENTRY POINTS

1

Training

- Ontology loading & standardization
- Reference alignment loading
- Training dataset construction
- Model fine-tuning

From standardized ontology

Offline Preprocessing

- Unified ontology representation
- Exact-match lookup structures
- Candidate retrieval indices
- Semantic representations (for hybrid retrieval)

2

3

Inference

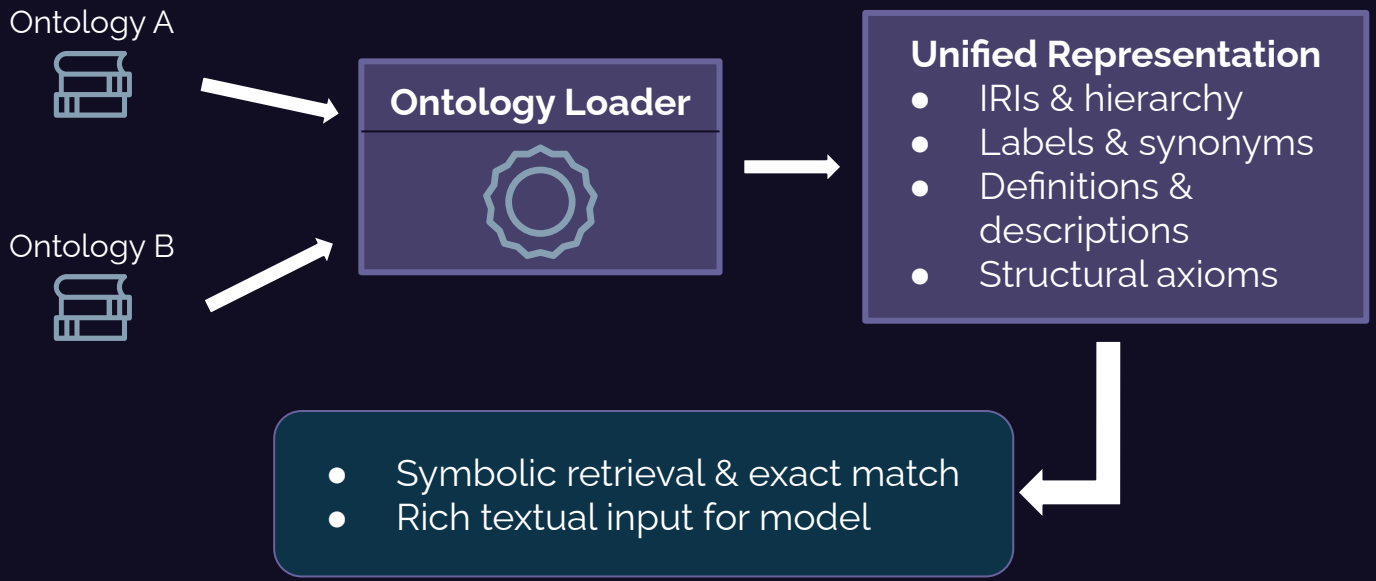
- Exact match shortcut
- Candidate selection
- Neural scoring

DATA ENGINEERING

04

ONTOLOGY LOADER FRAMEWORK

Load & standardize one or more ontologies into a unified representation



DATASET CONSTRUCTION

H3

Convert ontologies and reference alignments into a supervised dataset. Each sample is a concept–concept pair labeled as match or non-match.

A training sample can be:

POSITIVE = gold-standard alignment

NEGATIVE = unrelated concept pair

HARD NEGATIVE = semantically confusing non-match

TRAINING STRATEGY & MODEL SELECTION

H4

GOAL: study how Transformer models perform in semantic alignment

CHOICE: fine-tuning cross-encoder and bi-encoder architectures

TRADE-OFF: semantic accuracy vs computational scalability

OFFLINE BUNDLE & RETRIEVAL DECISIONS

H1

H2

The **offline bundle** is a compiled representation of the ontology: it is independent from incoming studies, built once per ontology version, and reused across all inference runs.

Exact Match Shortcut

- Constant-time lookup on labels & synonyms
- Maximal score, no model call

Candidate Selection

- Subword inverted index
- IDF-weighted lexical evidence
- Top-k candidates for neural scoring

- The ontology becomes an operational knowledge base, not something reprocessed at every inference step.

CANDIDATE SELECTION: FAILURE MODES & ROBUSTNESS

H2

Failure Modes

No lexical overlap

- Correct class not retrieved
- Never scored → alignment impossible

Lexical but wrong candidates

- Superficial similarity
- True match filtered out

Mitigation & Robustness

- Explicit detection of retrieval failures (no silent failures)
- Hybrid retrieval (lexical + semantic signals) (under investigation)
- Human-in-the-loop for unresolved cases (ontology evolution)

- Retrieval errors are structural, not model errors.
- Robustness must be addressed before neural scoring.

WORK MANAGEMENT STRUCTURE

05

WORK PACKAGES TIMELINE

		Weeks									
WP	OBJECTIVE	1	2	3	4	5	6	7	8	9	10
WP1	Domain understanding & conceptual design										
WP2	Data loading tools & training dataset construction										
WP3	SOTA analysis, baseline definition & model selection										
WP4	Model implementation, optimization & OAEI/MELT-ready output										
WP5	Final delivery, replicability & reporting										

WORK MANAGEMENT & NEXT STEPS

COMPLETED: system design, ontology loader framework, dataset construction pipeline

ONGOING: baseline definition and Transformer model training and comparison

NEXT: experimental validation on benchmarks and quantitative evaluation (precision, recall, F1)

THANK YOU

ANY QUESTIONS?