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 BACKGROUND

01

GIVING STRUCTURE TO THE VOICE OF PATIENTS

Repertorio transforms the *voice of the patient* into **STRUCTURED EVIDENCE**

Supporting:

- Clinical research
- Pharmaceutical R&D
- Market access strategies

THE CHALLENGES

DATA FRAGMENTATION

Critical healthcare evidence is widely scattered across isolated studies, creating disconnected data silos

SEMANTIC INCONSISTENCY

Different terms and formats limit the meaningful integration of heterogeneous data sources.

FROM PATIENT EVIDENCE TO BETTER DECISIONS

Repertorio generates **Quantified Evidence** to support regulators, pharma, and researchers in key decisions.

RISK-BENEFIT TRADE-OFFS

Optimizing clinical outcomes against safety profiles

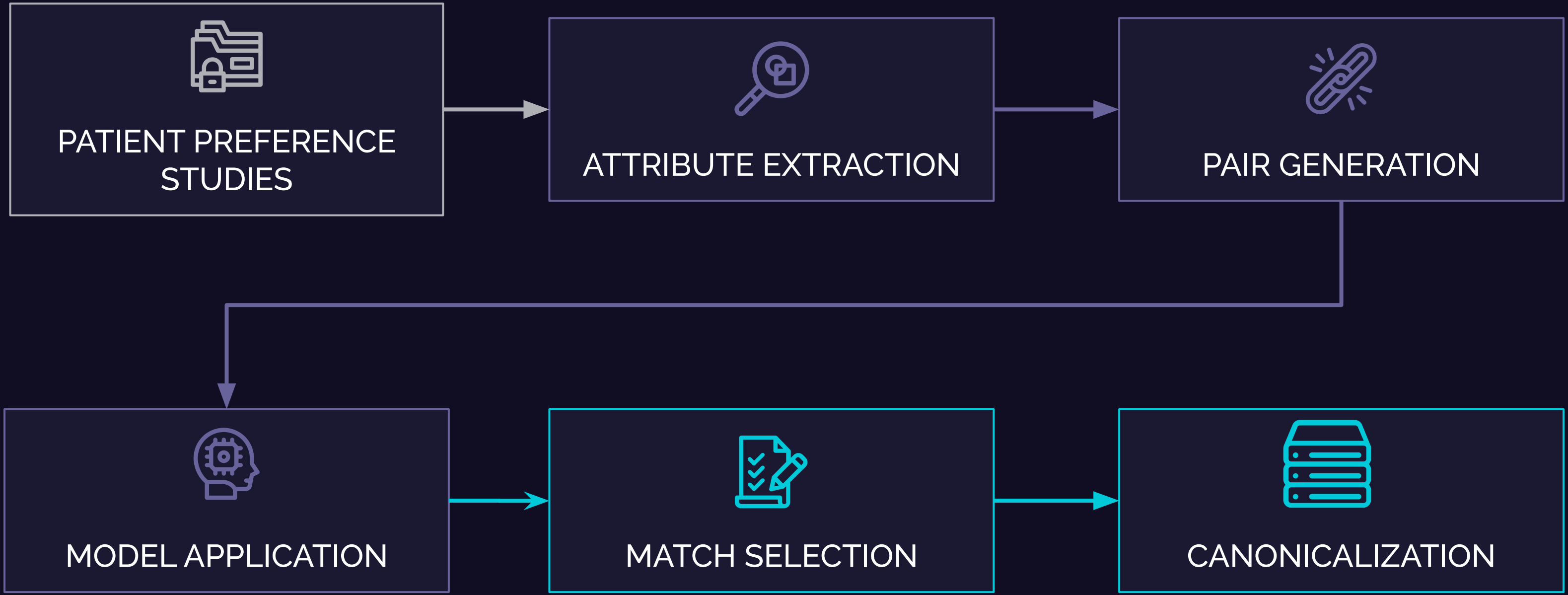
WILLINGNESS TO PAY

Quantifying value for market access

MAXIMUM ACCEPTABLE RISK

Defining safety boundaries for regulators

REPERTORIO FUNCTIONAL DIAGRAM



PATIENT PREFERENCE STUDIES

ATTRIBUTE EXTRACTION

PAIR GENERATION

MODEL APPLICATION

MATCH SELECTION

CANONICALIZATION

PROJECT VALUE PROPOSITION

02

THE VALUE PROPOSITION

Developing an **AI-POWERED SEMANTIC ENGINE** that automatically harmonizes fragmented data.

Enabling Repertorio to scale the analysis of patient preferences for faster, **EVIDENCE-BASED DECISIONS.**

3 GOOD HEALTH AND WELL-BEING

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

PROJECT GENERAL OBJECTIVE

03

GENERAL OBJECTIVE

BUILD THE AI ENGINE

Design and deploy the core AI model that aligns heterogeneous labels and concepts across ontologies

SEMANTIC MAPPING

Provide reliable similarity scores that Repertorio will use to map raw attributes to canonical concepts in their internal ontology

EXAMPLE

Fatigue ⇔ Asthenia ⇔ Lack of energy

THE RESEARCH OBJECTIVE

Developing a Context-Aware Semantic Alignment Engine

EVALUATE TRANSFORMER EFFICIENCY

Assess how models like BERT and S-BERT perform on complex biomedical data compared to traditional methods.

OVERCOME LEXICAL LIMITATIONS

Move beyond simple string matching to capture the actual *meaning* and *context* of clinical terms.

VALIDATE PERFORMANCE

Benchmark the new approach against established standards (OAEI) to ensure reliability.

THE EXPECTED OUTCOME

Reliable Semantic Foundation

Advanced NLP bridges terminology gaps, ensuring consistent, trustworthy data across the platform.

Advanced Analytics

Unlocked by consistent data, the platform will offer enhanced search and meta-analysis for deeper insights

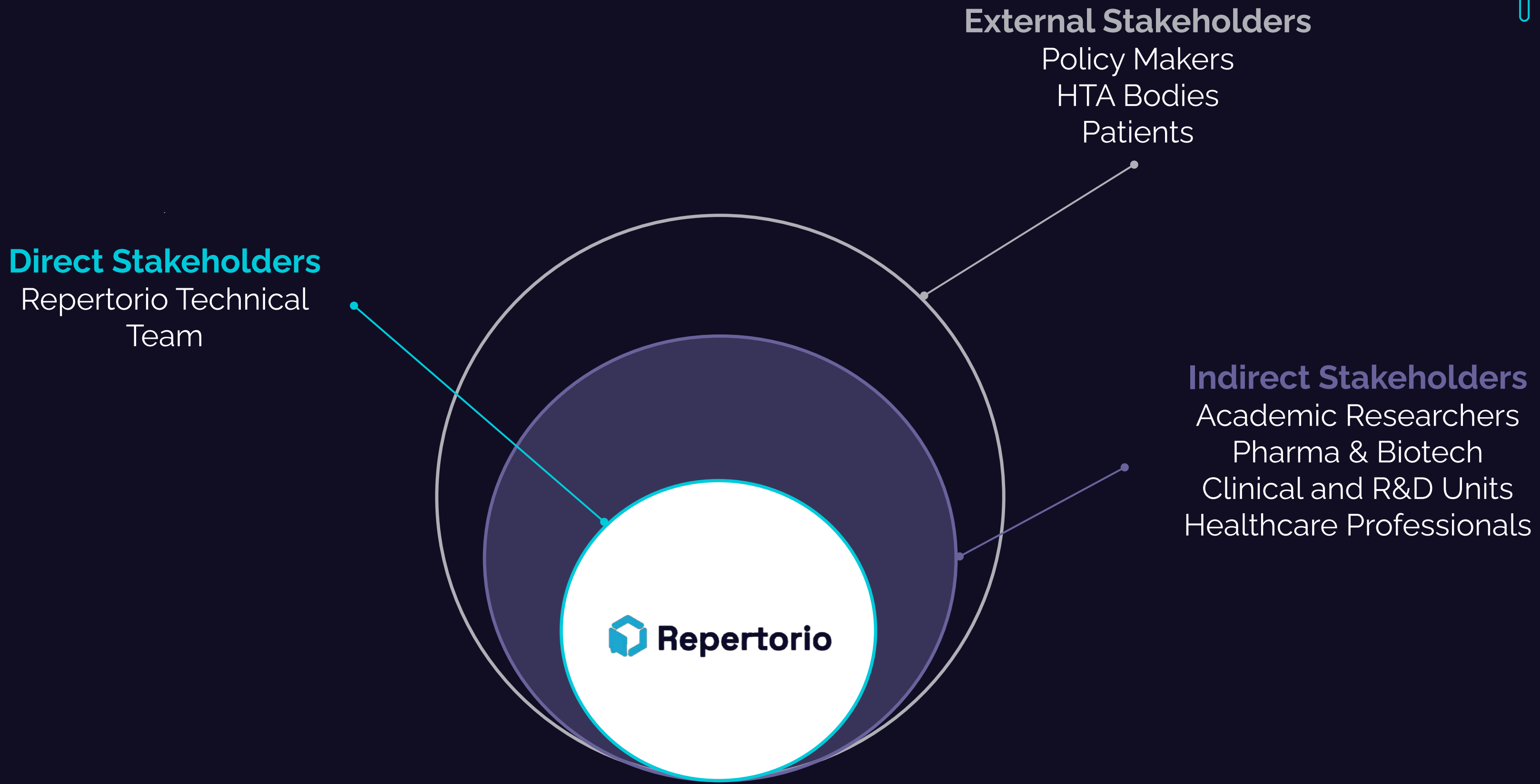
Stakeholder Value

These capabilities will deliver essential indirect benefits for researchers, clinicians, and ultimately, patients.

PROJECT DESIGN

04

THE STAKEHOLDERS



External Stakeholders
Policy Makers
HTA Bodies
Patients

Direct Stakeholders
Repertorio Technical Team

Indirect Stakeholders
Academic Researchers
Pharma & Biotech
Clinical and R&D Units
Healthcare Professionals

▶ THE PERSONA

DEMOGRAPHICS

45 years old • Turin, Italy
Senior Data Engineer

TASKS & HABITS

- Writes Python scripts daily.
- Aligns continuously with R&D.

PAIN POINTS

- Frustrated by brittle scripts.
- Manual data alignment is a bottleneck.

JOHN SMITH
LEAD DATA ENGINEER

NEEDS AND GOALS

- Automated, scalable alignment pipeline.
- Stop being the manual bottleneck.

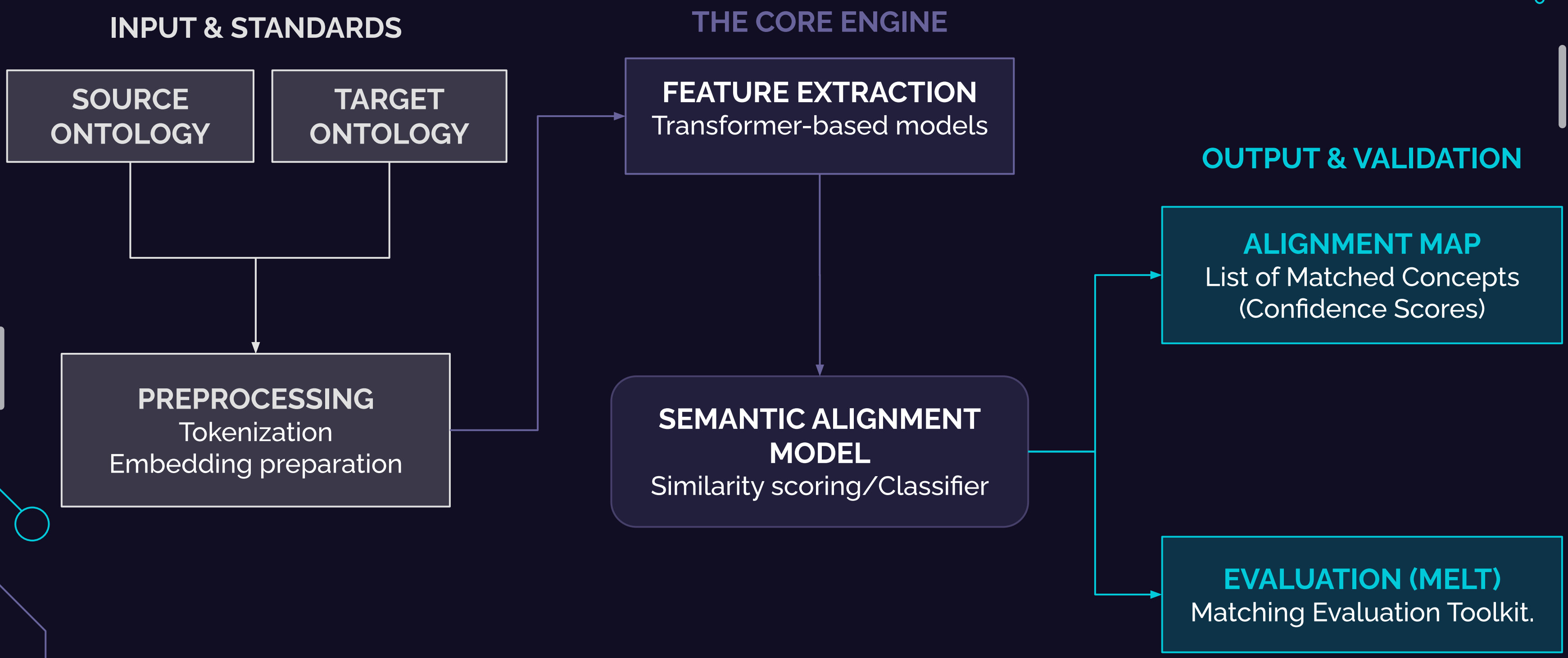
EXPECTED OUTPUT

- Reliable AI Model (API).
- Clear semantic match scores.

DATA USAGE

- Input: Messy OAEI / OWL files.
- Output: Clean Knowledge Graph.

SYSTEM FUNCTIONAL DIAGRAM



INPUT & STANDARDS

SOURCE ONTOLOGY

TARGET ONTOLOGY

PREPROCESSING
Tokenization
Embedding preparation

THE CORE ENGINE

FEATURE EXTRACTION
Transformer-based models

SEMANTIC ALIGNMENT MODEL
Similarity scoring/Classifier

OUTPUT & VALIDATION

ALIGNMENT MAP
List of Matched Concepts
(Confidence Scores)

EVALUATION (MELT)
Matching Evaluation Toolkit.

SYSTEM REQUIREMENTS

	FUNCTIONAL	NON-FUNCTIONAL
MUST HAVE	*Load*: the system must load an OAEI dataset (a track folder). *Build*: the system must automatically build a dataset of pairs (concept A, concept B) and their associated text/graph features, ready for training. *Train*: the system must allow training a binary classifier for ontology alignment. *Evaluate*: the system must be able to evaluate the model on a test set and produce standard metrics. *Generate*: the system must be able to generate an alignment file in one of the standard formats (e.g., Alignment API RDF) starting from the model's predictions	*Usability/Hardware Accessibility*: the system must be executable on machines with standard GPUs. *Reproducibility*: The system must be reproducible: same inputs, same parameters => same results.
SHOULD HAVE	*Compare*: the system should implement and compare semantic encoding using domain-specific Language Models (e.g., BioBERT, SciBERT) against generic models (e.g., RoBERTa).	*Transparency*: the system should be transparent regarding model version, OAEI tracks used for training, main hyperparameters.
COULD HAVE	*Inspect*: the Domain Expert could be able to view and inspect the proposed alignments, with clear visibility of ambiguous cases.	*Modularity*: the codebase could be structured into distinct modules, allowing the Repertorio team to retrain the model on future internal data.
WON'T HAVE	*Collect*: the system won't collect questionnaires or collect raw patient preference data from hospitals or clinics.	*Real-time inference*: the system won't require real-time processing.

PROJECT WORK MANAGEMENT STRUCTURE

05

THE WORK PACKAGES OVERVIEW

WP1 — Domain Understanding & High-Level Design

- **T1 ->** Stakeholder, persona & pipeline analysis
- **T2 ->** Requirements definition + High-Level Functional Diagram
- **T3 ->** Dataset inspection & MELT standards
- **T4 ->** Management Plan Construction

WP2 — Data Preparation & Dataset Builder

- **T1 ->** OAEI Loader (ontologies + reference alignment)
- **T2 ->** Dataset Builder (positive + negative samples)
- **T3 ->** Train/Val/Test split with leakage prevention

WP3 — Literature Review & Architecture Selection

- **T1 ->** Literature review
- **T2 ->** Baseline implementation (TF-IDF + cosine**)**
- **T3 ->** Comparative evaluation (precision/recall/F1)
- **T4 ->** Final architecture selection

WP4 — Model Training, Tuning & Evaluation

- **T1 ->** Model implementation & training pipeline
- **T2 ->** Hyperparameter tuning and error analysis
- **T3 ->** MELT-compatible alignment file generation

WP5 — Packaging & Final Presentation

- **T1 ->** Code packaging (modular repo: loader, builder, baseline, model)
- **T2 ->** Technical documentation + inference examples
- **T3 ->** Final report + slide deck (Checkpoint 3)
- **T4 ->** Public release of the full reproducible pipeline

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								■	■	■

THANK YOU

ANY QUESTIONS?