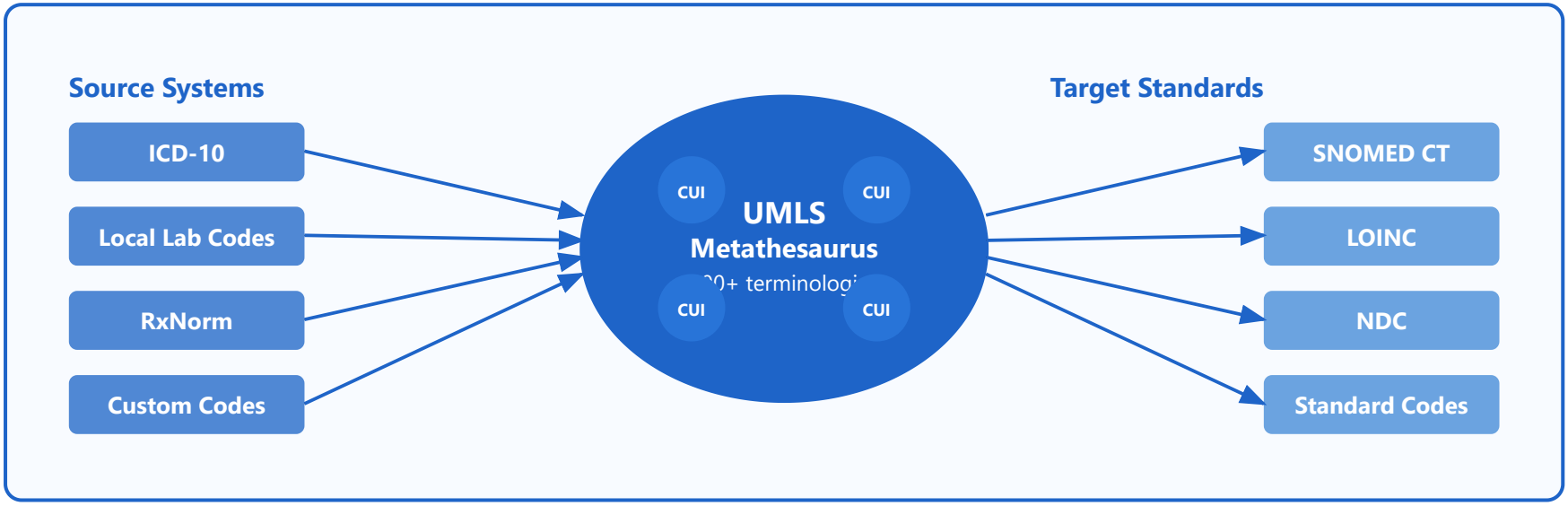
Ontology Mapping



Crosswalk Creation

- ICD-10 to SNOMED CT
- LOINC to local lab codes
- RxNorm to NDC
- Manual and automated approaches

Automated Mapping

- String similarity algorithms
- Lexical matching
- Machine learning classifiers
- Natural language processing

✓ Validation Methods

UMLS Metathesaurus

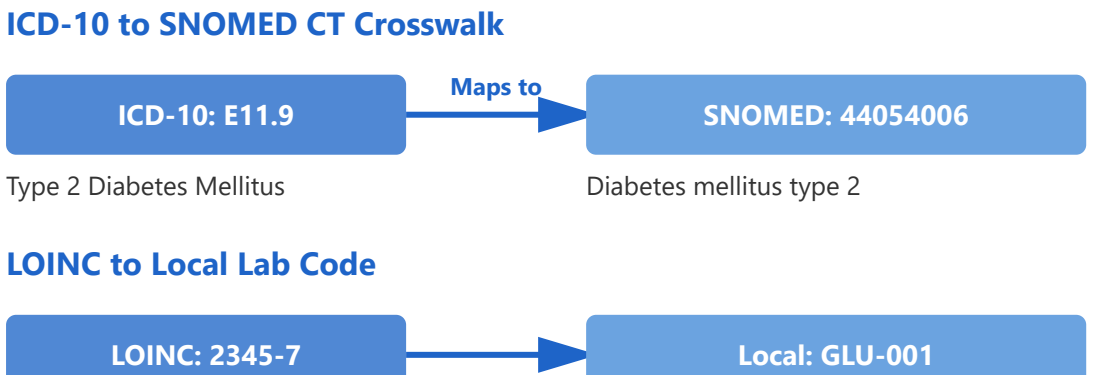
- Expert review
- Dual coding
- Inter-rater reliability
- Continuous quality improvement

- Unified Medical Language System
- Integrates 200+ terminologies
- Concept Unique Identifiers (CUI)
- Relationship mappings across systems

1. Crosswalk Creation: Detailed Overview

What is a Crosswalk?

A crosswalk (or mapping table) is a systematic translation guide that establishes relationships between concepts in different coding systems. It serves as a bridge that enables data exchange and interoperability between disparate healthcare information systems.



Types of Crosswalk Relationships

Relationship Type	Description	Example
One-to-One	Direct equivalent mapping between codes	ICD-10 I10 ↔ SNOMED 38341003 (Hypertension)
One-to-Many	Single source code maps to multiple target codes	ICD-10 E11 → Multiple SNOMED codes for diabetes subtypes
Many-to-One	Multiple source codes map to single target code	Multiple local codes → Single LOINC code
Partial/Approximate	No exact equivalent, closest match used	Legacy system code ≈ Modern standard code

💡 Real-World Example: RxNorm to NDC Mapping

Scenario: A hospital needs to map prescription data from their system using RxNorm codes to billing codes in NDC format.

```
RxNorm Code: 213269 (Ibuprofen 200 MG Oral Tablet) ↓ Maps to Multiple NDCs: ├─ NDC: 50580-0608-01 (Generic manufacturer A) ├─ NDC: 00904-5816-60 (Generic manufacturer B) └─ NDC: 41250-0780-10 (Generic manufacturer C)
```

Challenge: One RxNorm code can map to dozens or hundreds of NDC codes because NDC is package-specific while RxNorm is ingredient-specific.

Crosswalk Creation Methods

- ✓ **Manual Expert Mapping:** Clinical experts review and establish mappings based on semantic equivalence
- ✓ **Algorithm-Assisted:** Software suggests potential mappings based on text similarity, which experts then validate
- ✓ **Published Crosswalks:** Use pre-existing mappings from authoritative sources (CMS, NLM, WHO)
- ✓ **Hybrid Approach:** Combine automated suggestions with expert validation and published resources

Challenges in Crosswalk Creation

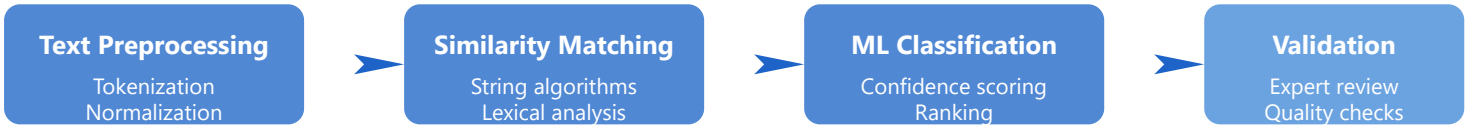
- ⚠ **Semantic Heterogeneity:** Different systems may define concepts at different levels of granularity
- ⚠ **Missing Equivalents:** Not all concepts in one system have direct counterparts in another
- ⚠ **Version Management:** Terminologies evolve; crosswalks must be updated when source or target systems change
- ⚠ **Context Dependency:** The appropriate mapping may depend on clinical context or use case

2. Automated Mapping: Computational Approaches

Overview of Automated Mapping

Automated mapping leverages computational algorithms to identify potential mappings between terminology systems. These methods significantly reduce manual effort while maintaining acceptable accuracy levels when properly validated.

Automated Mapping Pipeline



Key Techniques:

- Levenshtein Distance: Measures character-level edit distance
- Cosine Similarity: Compares term vectors in semantic space
- Neural Embeddings: Deep learning models (BERT, BioBERT) for semantic similarity

1. String Similarity Algorithms

💡 Levenshtein Distance Example

Measures the minimum number of single-character edits (insertions, deletions, substitutions) needed to transform one string into another.

```
Source Term: "myocardial infarction" Target Term: "myocardial infarction acute"
Levenshtein Distance: 7 (adding " acute") Similarity Score: 77% (normalized)
```

```
Interpretation: High similarity suggests potential mapping
```

Algorithm	Strength	Best Use Case
Levenshtein Distance	Handles misspellings and variations	Similar terms with minor differences
Jaro-Winkler	Weights beginning of strings more	Names and medical terms with prefixes
N-gram Matching	Robust to word order changes	Multi-word medical phrases
Soundex/Metaphone	Phonetic matching	Terms with varied spellings

2. Lexical Matching Techniques

Lexical matching analyzes the words and linguistic structure of concept descriptions to identify semantic relationships.

- ✓ **Token Overlap:** Counts shared words between descriptions
- ✓ **Synonym Expansion:** Uses dictionaries to match synonymous terms
- ✓ **Stopword Removal:** Eliminates common words to focus on meaningful terms
- ✓ **Stemming/Lemmatization:** Reduces words to root forms (e.g., "running" → "run")

3. Machine Learning Classifiers

Modern ML approaches learn mapping patterns from training data and can predict mappings for new concept pairs.

ML-Based Mapping Workflow

```
Training Phase: 1. Collect validated mapping pairs (source, target) 2. Extract features: - String similarity scores - Lexical overlap metrics - Semantic embeddings - Hierarchical relationships 3. Train classifier (Random Forest, Neural Network, etc.) 4. Evaluate on test set Prediction Phase: 1. Input: unmapped source concept 2. Generate candidate targets 3. Extract features for each candidate 4. Classifier predicts match probability 5. Rank candidates by confidence score 6. Present top candidates for expert review
```

4. Natural Language Processing (NLP)

Advanced NLP techniques leverage deep learning and contextual understanding to improve mapping accuracy.

NLP Technique	Application in Mapping	Example Tool
Word Embeddings	Capture semantic relationships in vector space	Word2Vec, GloVe
Transformer Models	Contextual understanding of medical terms	BERT, BioBERT, ClinicalBERT
Named Entity Recognition	Identify medical concepts in text	scispaCy, MetaMap

NLP Technique	Application in Mapping	Example Tool
Semantic Similarity	Measure conceptual closeness	Sentence-BERT, Universal Sentence Encoder

💡 BioBERT Mapping Example

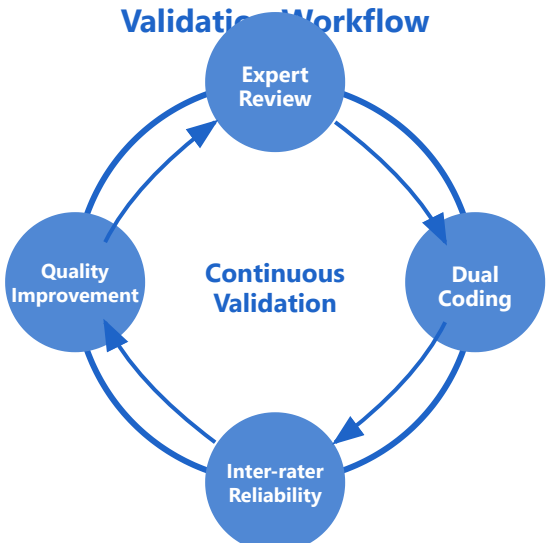
BioBERT, trained on biomedical literature, can identify semantically similar concepts even when terminology differs:

```
Source: "heart attack" (lay term) Target candidates: 1. "myocardial infarction" - Similarity: 0.94 ✓ High match 2. "cardiac arrest" - Similarity: 0.72 3. "angina pectoris" - Similarity: 0.68 The model learned from millions of medical texts that "heart attack" and "myocardial infarction" are synonymous, despite no lexical overlap.
```

✓ 3. Validation Methods: Ensuring Mapping Quality

Why Validation is Critical

Even with sophisticated automated mapping, validation is essential to ensure clinical accuracy, patient safety, and regulatory compliance. Incorrect mappings can lead to misdiagnosis, billing errors, and compromised data analytics.



1. Expert Review

Clinical domain experts (physicians, nurses, terminology specialists) manually review and validate mappings to ensure clinical accuracy and appropriateness.

💡 Expert Review Process

```
Step 1: Present mapping candidate to expert Source: ICD-10 E10.9 (Type 1 diabetes without complications) Target: SNOMED 46635009 (Diabetes mellitus type 1) Automated Confidence: 0.89 Step 2: Expert evaluation criteria ✓ Semantic equivalence: Do concepts mean the same thing? ✓ Clinical context: Appropriate in all care settings? ✓ Granularity match: Same level of detail? ✓ Usage consistency: Aligned with clinical practice? Step 3: Expert decision [Approve] [Reject] [Modify] [Flag for discussion] Step 4: Documentation - Record rationale for decision - Note any contextual limitations - Suggest alternative mappings if applicable
```

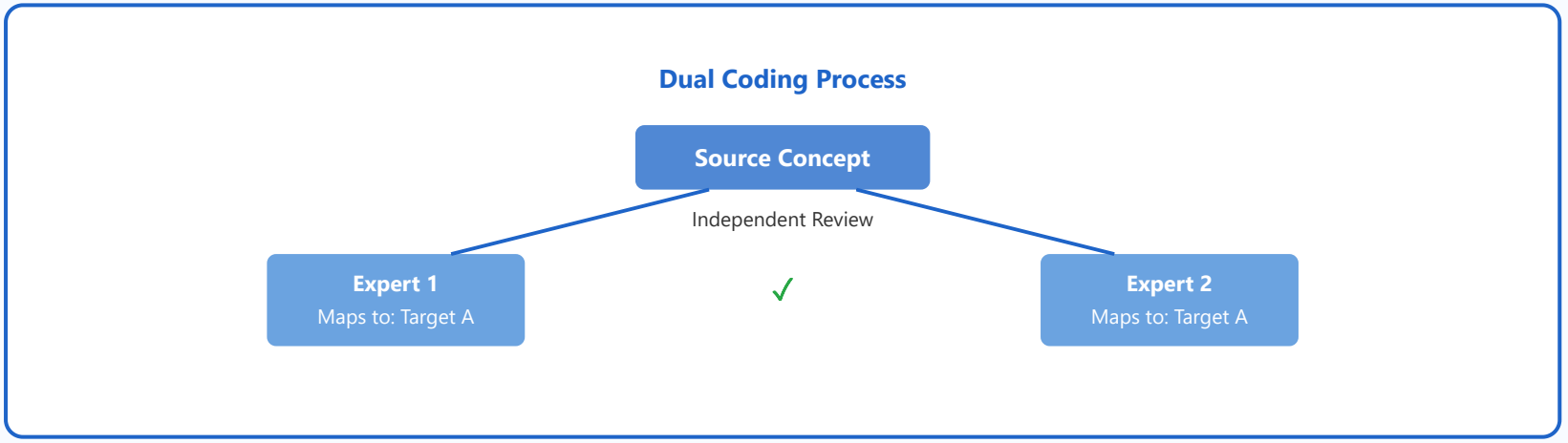
✓ **Advantages:** High accuracy, catches subtle semantic issues, incorporates clinical judgment

✓ **Best for:** Complex cases, ambiguous mappings, high-risk clinical domains

⚠ **Limitations:** Time-consuming, expensive, subject to individual bias, doesn't scale well

2. Dual Coding (Independent Review)

Two or more experts independently review and code the same set of concepts. Discrepancies are identified and resolved through discussion or adjudication.



💡 Dual Coding Scenario

Scenario: Mapping 100 local procedure codes to CPT codes

```
Expert 1 and Expert 2 independently map all 100 codes Results: - Agreement on 87 codes
(87% initial agreement) - Disagreement on 13 codes (13% discrepancy rate) Discrepancy
Resolution: Local Code LP-2045 (Endoscopic examination of stomach) Expert 1 → CPT 43235
(Upper GI endoscopy, diagnostic) Expert 2 → CPT 43239 (Upper GI endoscopy with biopsy)
```

```
Resolution process: 1. Review case details and clinical context 2. Consult additional
expert or adjudicator 3. Examine code definitions and guidelines 4. Reach consensus:
Use 43235 for diagnostic; create additional rule for 43239 if biopsy documented
```

3. Inter-rater Reliability (IRR)

Statistical measure of agreement between multiple coders. Common metrics include Cohen's Kappa, Fleiss' Kappa, and percent agreement.

Kappa Value	Level of Agreement	Interpretation
< 0.00	Poor	Agreement no better than chance
0.00 - 0.20	Slight	Minimal agreement
0.21 - 0.40	Fair	Modest agreement
0.41 - 0.60	Moderate	Reasonable agreement
0.61 - 0.80	Substantial	High agreement
0.81 - 1.00	Almost Perfect	Near-complete agreement

IRR Calculation Example

```
Study: 3 experts mapping 50 diagnosis codes Agreement matrix: Expert 2 Agree Disagree
Expert 1 Agree 42 3 Disagree 2 3 Observed Agreement (Po) = (42 + 3) / 50 = 0.90
Expected Agreement (Pe) = 0.82 (by chance) Cohen's Kappa = (Po - Pe) / (1 - Pe) = (0.90
```

```
- 0.82) / (1 - 0.82) = 0.44 (Moderate agreement) Interpretation: Agreement is better
than chance, but training needed to improve consistency
```

4. Continuous Quality Improvement (CQI)

Ongoing process of monitoring mapping quality, identifying issues, and implementing improvements over time.

- ✓ **Regular Audits:** Periodic sampling and review of mappings to detect errors or drift
- ✓ **Feedback Loops:** Users report problematic mappings; issues tracked and resolved
- ✓ **Performance Metrics:** Track accuracy, coverage, and consistency over time
- ✓ **Version Control:** Maintain mapping history and rationale for changes
- ✓ **Training Programs:** Regular education for coders on updated standards and best practices

💡 CQI Dashboard Metrics

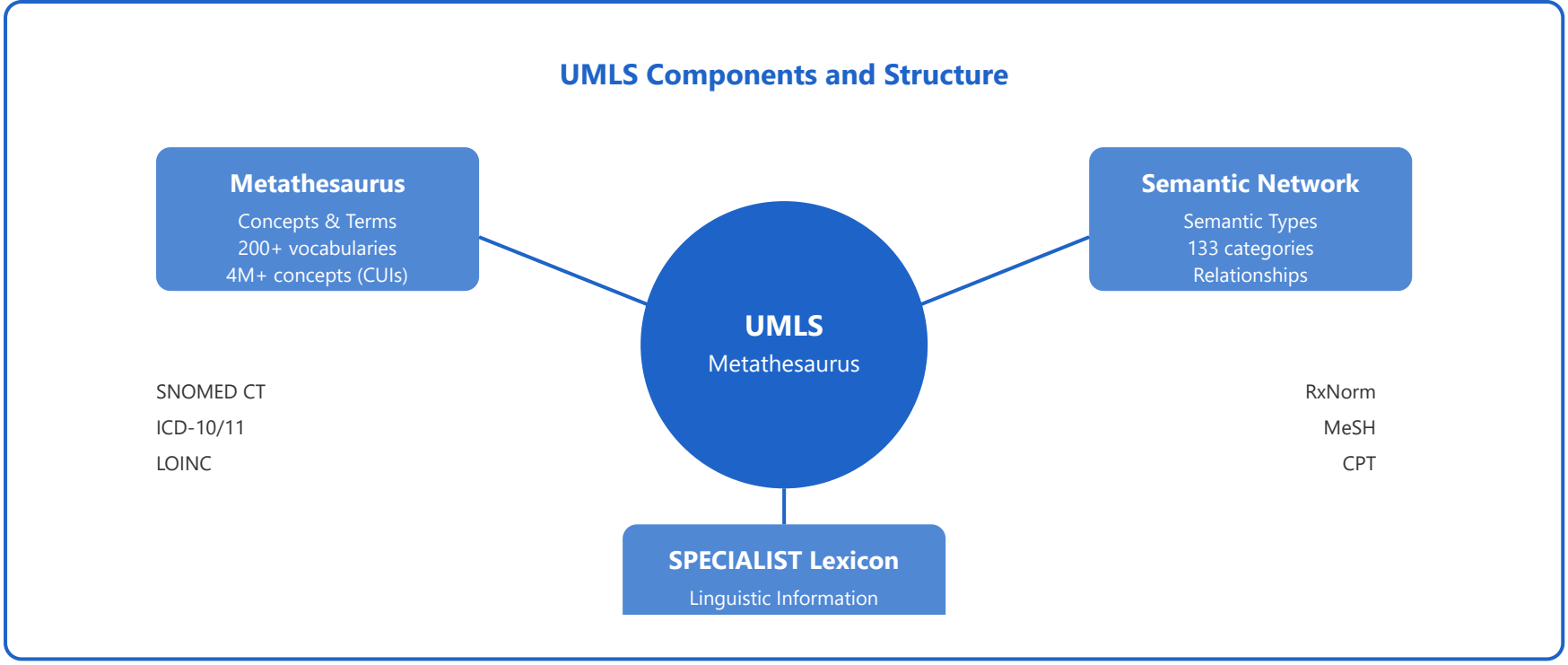
```
Monthly Mapping Quality Report - October 2024 Coverage: Total Source Codes: 15,842
Successfully Mapped: 15,201 (95.9%) Unmapped: 641 (4.1%) Accuracy (from random audit of
500 mappings): Correct: 478 (95.6%) Incorrect: 15 (3.0%) Questionable: 7 (1.4%) Common
Issues Identified: 1. Granularity mismatch (45% of errors) 2. Outdated terminology (30%
of errors) 3. Ambiguous source descriptions (25% of errors) Action Items: ✓ Update 23
```

```
mappings affected by recent code changes ✓ Provide additional training on granularity
matching ✓ Request clarification on 47 ambiguous source codes
```

🌐 4. UMLS Metathesaurus: The Universal Hub

What is UMLS?

The Unified Medical Language System (UMLS) is a comprehensive knowledge source developed by the U.S. National Library of Medicine (NLM). It integrates over 200 biomedical vocabularies and standards, providing a unified framework for mapping between different terminology systems.



Core Components of UMLS

1. Metathesaurus: The heart of UMLS, containing concepts and their relationships across terminology systems.

Element	Description	Example
CUI (Concept Unique Identifier)	Unique identifier for each concept	C0011847
AUI (Atom Unique Identifier)	Unique identifier for each term variant	A8345263
String	Textual representation of concept	"Diabetes Mellitus"
Source	Origin terminology system	SNOMED CT, ICD-10, etc.
Semantic Type	Broad category of concept	Disease or Syndrome

2. Semantic Network: Provides high-level categories and relationships between concepts.

3. SPECIALIST Lexicon: Contains linguistic information and tools for natural language processing of biomedical text.

💡 How UMLS Connects Multiple Terminologies

Example: The concept "Type 2 Diabetes Mellitus" exists in multiple terminologies with different codes. UMLS unifies them under a single CUI.

```
UMLS CUI: C0011860 (Type 2 Diabetes Mellitus) Connected Terms from Different Sources: ├─
SNOMED CT: 44054006 │ └─ "Diabetes mellitus type 2 (disorder)" ├─ ICD-10-CM: E11 │ └─ "Type 2
diabetes mellitus" ├─ ICD-9-CM: 250.00 │ └─ "Diabetes mellitus without mention of
complication" ├─ MeSH: D003924 │ └─ "Diabetes Mellitus, Type 2" ├─ MEDCIN: 33716 │ └─
"diabetes mellitus type 2" └─ NCI Thesaurus: C26747 └─ "Diabetes Mellitus, Non-Insulin-
Dependent" Semantic Type: Disease or Syndrome (T047) Relationships: - ISA: Diabetes Mellitus
(C0011849) - associated_with: Insulin Resistance (C0021655) - may_be_treated_by: Metformin
(C0025598)
```

Using UMLS for Ontology Mapping

- ✓ **Cross-System Translation:** Map from any source system to any target system via shared CUIs
- ✓ **Semantic Enrichment:** Add semantic type information and relationships to enhance mapping
- ✓ **Synonym Finding:** Discover alternative terms and synonyms across languages and systems
- ✓ **Hierarchical Navigation:** Traverse concept hierarchies to find parent/child concepts

✓ **Quality Assurance:** Leverage expert-curated mappings maintained by NLM

💡 UMLS API Usage Example

Practical example of using UMLS to map between ICD-10 and SNOMED CT:

```
Query: Map ICD-10 code "I10" (Hypertension) to SNOMED CT Step 1: Find CUI for ICD-10 I10
Request: GET /rest/search/current?string=I10&sabs=ICD10CM Response: CUI = C0020538 Step 2: Get
concept details Request: GET /rest/content/current/CUI/C0020538 Response: - Preferred name:
"Hypertensive disease" - Semantic Type: T047 (Disease or Syndrome) Step 3: Find SNOMED CT code
for same CUI Request: GET /rest/content/current/CUI/C0020538/atoms?sabs=SNOMEDCT_US Response:
- SNOMED CT: 38341003 - Term: "Hypertensive disorder, systemic arterial (disorder)" Result:
ICD-10 I10 → UMLS C0020538 → SNOMED CT 38341003
```

UMLS Statistics and Coverage

Metric	Current Value (2024)
Source Vocabularies	200+
Unique Concepts (CUIs)	~4.5 million
Total Terms/Atoms	~15 million
Languages Supported	25+
Semantic Types	133

Metric	Current Value (2024)
Semantic Relationships	54

Key Integrated Terminologies in UMLS

UMLS includes comprehensive coverage of major healthcare terminologies:

- ✓ **Clinical:** SNOMED CT, ICD-10/11, ICD-9-CM, DSM-5, ICPC
- ✓ **Laboratory:** LOINC (Logical Observation Identifiers Names and Codes)
- ✓ **Medications:** RxNorm, NDC (National Drug Code), ATC
- ✓ **Procedures:** CPT (Current Procedural Terminology), HCPCS
- ✓ **Research:** MeSH (Medical Subject Headings), NCI Thesaurus
- ✓ **Nursing:** NANDA-I, NIC (Nursing Interventions Classification), NOC

💡 Real-World UMLS Application

Use Case: A research institution needs to integrate patient data from multiple hospitals using different coding systems for a diabetes study.

```
Challenge: Hospital A: Uses ICD-10 for diagnoses Hospital B: Uses SNOMED CT for clinical documentation Hospital C: Uses local codes mapped to ICD-9 Solution Using UMLS: 1. Extract all diabetes-related codes from each hospital 2. Map each code to UMLS CUI: Hospital A: ICD-10 E11.x → CUI C0011860 Hospital B: SNOMED 44054006 → CUI C0011860 Hospital C: Local code → ICD-9 250.00 → CUI C0011860 3. Use CUI as common identifier for integration 4. Query UMLS for related concepts: - Retrieve all complications (via relationships) - Find associated medications (via "may_treat" relations) - Identify relevant lab tests (via semantic associations) Outcome: ✓ Unified dataset with 15,000 patients ✓ Consistent concept identification across sites ✓ Enriched data with semantic relationships ✓ Enabled comprehensive diabetes outcomes analysis
```

Challenges and Limitations of UMLS

⚠ **Complexity:** Large size and complexity can be overwhelming for new users

⚠ **Licensing:** Some source vocabularies have usage restrictions despite being in UMLS

⚠ **Maintenance:** Requires effort to stay current with updates to source terminologies

⚠ **Ambiguity:** Some concepts may have multiple potential mappings depending on context

⚠ **Coverage Gaps:** Not all specialized or emerging terminologies are included

Access and Tools

UMLS is freely available after obtaining a license from the National Library of Medicine. Various tools and resources support UMLS usage:

✓ **MetamorphoSys:** Installation and customization tool for local UMLS deployment

- ✓ **UTS (UMLS Terminology Services):** Web-based browser and REST API for queries
- ✓ **MetaMap:** NLP tool for mapping free text to UMLS concepts
- ✓ **UMLS Knowledge Sources:** Downloadable files in various formats (RRF, SQL, etc.)

Key Takeaways

- ✓ **Crosswalks** enable systematic translation between terminology systems, essential for data integration and interoperability
- ✓ **Automated mapping** combines string similarity, lexical analysis, machine learning, and NLP to accelerate mapping while reducing manual effort
- ✓ **Validation methods** ensure mapping quality through expert review, dual coding, inter-rater reliability, and continuous improvement
- ✓ **UMLS Metathesaurus** serves as a comprehensive hub connecting 200+ terminologies, facilitating mappings across the entire healthcare ecosystem

✓ **Successful ontology mapping** requires a balanced approach: leveraging automation for efficiency while maintaining rigorous validation for accuracy and clinical safety