VDLB 2026 Tutorial

Interoperability in Healthcare: A Primer and New Frontiers

KTH VETENSKAP OCH KONST

NYU

TU WIEN

Who are we?

Andra Ionescu

KTH Royal Institute of Technology, *Sweden*

Who are we?

Andra Ionescu

KTH Royal Institute of Technology, *Sweden*

Paris Carbone

KTH Royal Institute of Technology, *Sweden*

Sebastiaan Meijer

KTH Royal Institute of Technology, *Sweden*

Jayanth Raghothama

KTH Royal Institute of Technology, *Sweden*

Who are we?

Andra Ionescu

KTH Royal Institute of Technology, *Sweden*

Paris Carbone

KTH Royal Institute of Technology, *Sweden*

Sebastiaan Meijer

KTH Royal Institute of Technology, *Sweden*

Christos Koutras

New York University, *USA*

Juliana Freire

New York University, *USA*

Jayanth Raghothama

KTH Royal Institute of Technology, *Sweden*

Who are we?

Andra Ionescu

KTH Royal Institute of Technology, *Sweden*

Paris Carbone

KTH Royal Institute of Technology, *Sweden*

Sebastiaan Meijer

KTH Royal Institute of Technology, *Sweden*

Christos Koutras

New York University, *USA*

Juliana Freire

New York University, *USA*

Katja Hose

TU Wien, *Austria*

Jayanth Raghothama

KTH Royal Institute of Technology, *Sweden*

Agenda

Scan for slides

https://tutorial.mediverse.se/

Agenda

01 Intro & First Generation

Motivation, Interoperability, How it started

Scan for slides

https://tutorial.mediverse.se/

Agenda

01 **Intro & First Generation**

Motivation, Interoperability, How it started

02 **Second Generation**

Standards and Challenges

Scan for slides

https://tutorial.mediverse.se/

Agenda

01 **Intro & First Generation**

Motivation, Interoperability, How it started

02 **Second Generation**

Standards and Challenges

03 **Third Generation**

LLMs, Graphs, Interactive Harmonisation

Scan for slides

https://tutorial.mediverse.se/

Agenda

01 **Intro & First Generation**
Motivation, Interoperability, How it started

02 **Second Generation**
Standards and Challenges

03 **Third Generation**
LLMs, Graphs, Interactive Harmonisation

04 **The New Frontier**
Open Questions and Future Work

Scan for slides

https://tutorial.mediverse.se/

PART 01

Motivation

Andra Ionescu, Jayanth Raghothama

Why now?

Why now?

PAST TUTORIAL · VLDB 2012

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability
- Integrating Healthcare Enterprise (IHE) profiles

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability
- Integrating Healthcare Enterprise (IHE) profiles
- Deployment: Turkey's National EHR system

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability
- Integrating Healthcare Enterprise (IHE) profiles
- Deployment: Turkey's National EHR system

THIS TUTORIAL

Here’s what we want to achieve with this tutorial:

GOALS FOR THIS TUTORIAL

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability
- Integrating Healthcare Enterprise (IHE) profiles
- Deployment: Turkey's National EHR system

THIS TUTORIAL

Here’s what we want to achieve with this tutorial:

GOALS FOR THIS TUTORIAL

1

Map today's standards landscape

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability
- Integrating Healthcare Enterprise (IHE) profiles
- Deployment: Turkey's National EHR system

THIS TUTORIAL

Here’s what we want to achieve with this tutorial:

GOALS FOR THIS TUTORIAL

1

Map today's standards landscape

2

Share the HI challenges and DM solutions

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability
- Integrating Healthcare Enterprise (IHE) profiles
- Deployment: Turkey's National EHR system

THIS TUTORIAL

Here’s what we want to achieve with this tutorial:

GOALS FOR THIS TUTORIAL

1

Map today's standards landscape

2

Share the HI challenges and DM solutions

3

Share real deployments & lessons

Why now?

PAST TUTORIAL · VLDB 2012

14 years since we last talked about it

Interoperability in healthcare has developed considerably since

WHAT IT COVERED

- EHR, PHR & device interoperability
- Integrating Healthcare Enterprise (IHE) profiles
- Deployment: Turkey's National EHR system

THIS TUTORIAL

Here’s what we want to achieve with this tutorial:

GOALS FOR THIS TUTORIAL

1

Map today's standards landscape

2

Share the HI challenges and DM solutions

3

Share real deployments & lessons

4

Chart the future direction

Dogac, Asuman. "Interoperability in eHealth systems." *Proceedings of the VLDB Endowment* 5.12 (2012): 2026-2027.

Interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged.

IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries
https://ieeexplore.ieee.org/document/182763

ONE DOES NOT SIMPLY
DEFINE INTEROPERABILITY

Interoperability

The Healthcare Information and Management Systems Society (HIMSS) defines four levels of interoperability

Technical

Establishes the inter-connectivity requirements to exchange data

Syntactic

Concerns the format and structure in which data is encoded and exchanged

Semantic

Concerns the ability to share meaning across systems

Organizational

Includes governance, policy, social, legal and organizational considerations

Interoperability

Key Challenge:

The epistemology of clinical practice is not the same as the information (domain) model that captures it in data systems.

In other words:

No agreement on the semantics (common domain models)

Which leads to:

No base ontology of epistemic entities (observations, etc.) which can be instantiated as an information model

Technical: Solved

Syntactic: Solved

Semantic: Addressed through standards

Organizational: Not even close

Semantic Interoperability

- For complex domains, semantics are key:
  - The information model that captures an episode of a blood pressure reading of a patient is *not* the same as the ontological description of blood pressure
- The *meaning* of data needs to be unambiguously exchanged, and *understood* by the receiver (system and user), *without* special measures
- Interoperability (semantic) needs to be an emergent phenomenon of underlying common domain models, not ad-hoc reengineering of endless specifications (current scenario)

Ref: Stenzhorn et.al. 2008

Interoperability

Secondary Challenges:

1. Legacy systems
2. Proliferation of vocabularies and terminologies
3. Poor adherence to standards and terminologies
4. Closed and proprietary information models
5. Variance in clinical practice, cultures and demographics (perhaps not something to be "solved")

Technical: Solved

Syntactic: Solved

Semantic: Addressed through standards

Organizational: Not even close

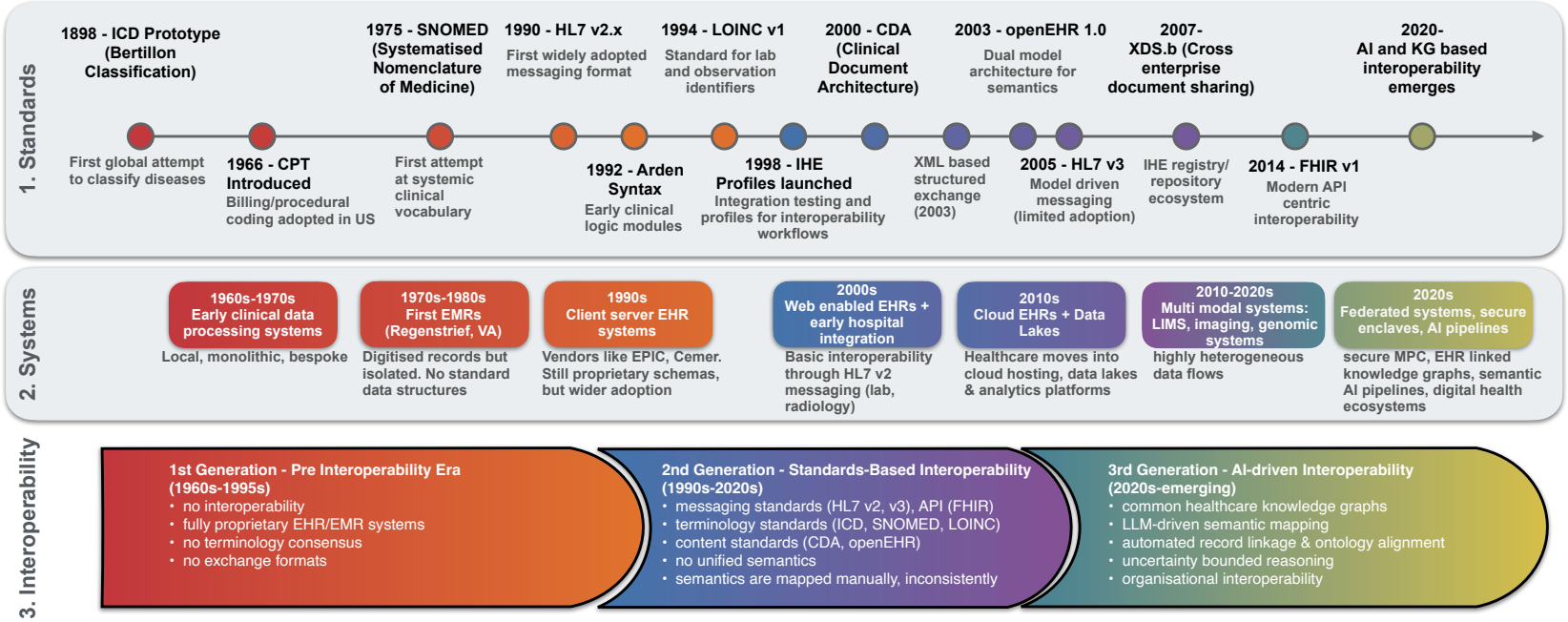
Ref: Beale et.al 2022

PART 01

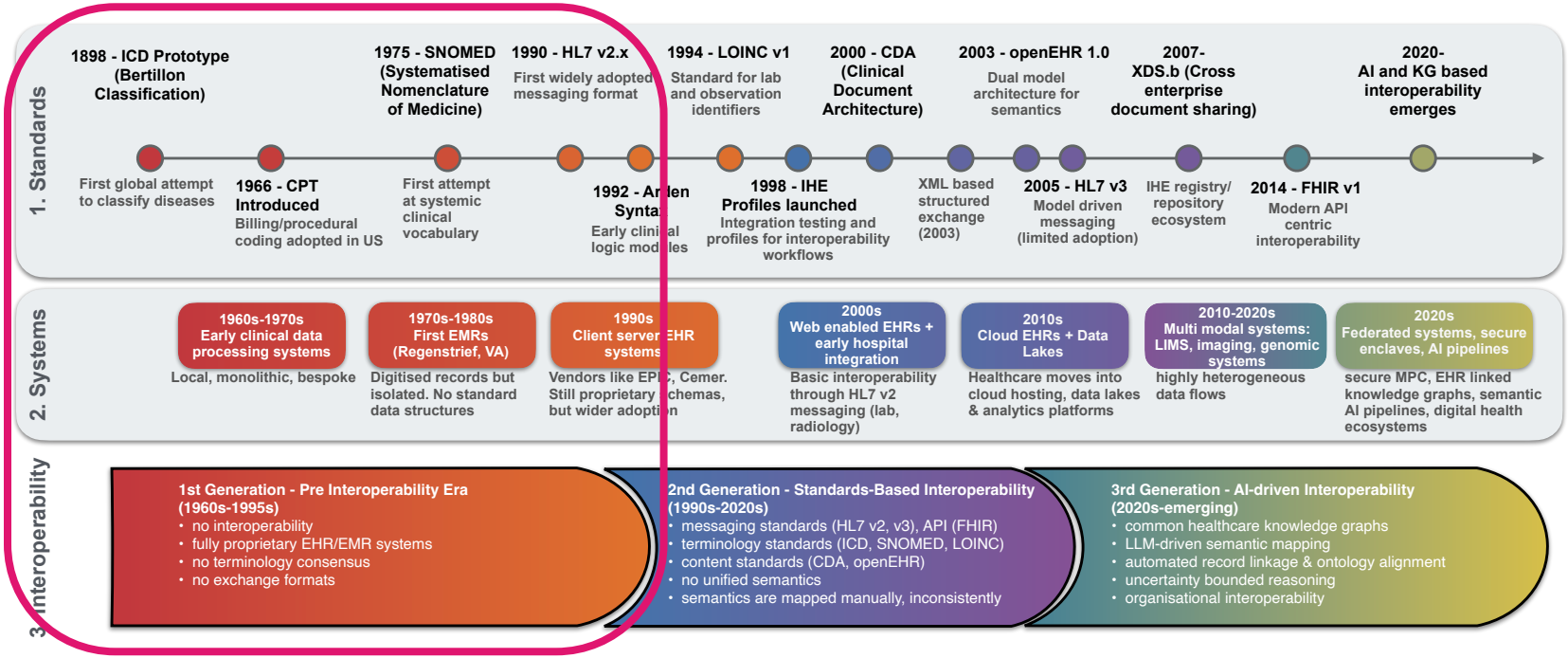
First Generation

Jayanth Raghothama

History


1. Standards
1898 - ICD Prototype (Bertillon Classification)
First global attempt to classify diseases
1966 - CPT Introduced
Billing/procedural coding adopted in US
1975 - SNOMED (Systematised Nomenclature of Medicine)
First attempt at systemic clinical vocabulary
1990 - HL7 v2.x
First widely adopted messaging format
1992 - Arden Syntax
Early clinical logic modules
1994 - LOINC v1
Standard for lab and observation identifiers
1998 - IHE Profiles launched
Integration testing and profiles for interoperability workflows
2000 - CDA (Clinical Document Architecture)
XML based structured exchange (2003)
2003 - openEHR 1.0
Dual model architecture for semantics
2005 - HL7 v3
Model driven messaging (limited adoption)
2007- XDS.b (Cross enterprise document sharing)
IHE registry/ repository ecosystem
2014 - FHIR v1
Modern API centric interoperability
2020- AI and KG based interoperability emerges
2. Systems
1960s-1970s Early clinical data processing systems
Local, monolithic, bespoke
1970s-1980s First EMRs (Regenstrief, VA)
Digitised records but isolated. No standard data structures
1990s Client server EHR systems
Vendors like EPIC, Cemer. Still proprietary schemas, but wider adoption
2000s Web enabled EHRs + early hospital integration
Basic interoperability through HL7 v2 messaging (lab, radiology)
2010s Cloud EHRs + Data Lakes
Healthcare moves into cloud hosting, data lakes & analytics platforms
2010-2020s Multi modal systems: LIMS, imaging, genomic systems
highly heterogeneous data flows
2020s Federated systems, secure enclaves, AI pipelines
secure MPC, EHR linked knowledge graphs, semantic AI pipelines, digital health ecosystems
3. Interoperability
1st Generation - Pre Interoperability Era (1960s-1995s)
• no interoperability
• fully proprietary EHR/EMR systems
• no terminology consensus
• no exchange formats
2nd Generation - Standards-Based Interoperability (1990s-2020s)
• messaging standards (HL7 v2, v3), API (FHIR)
• terminology standards (ICD, SNOMED, LOINC)
• content standards (CDA, openEHR)
• no unified semantics
• semantics are mapped manually, inconsistently
3rd Generation - AI-driven Interoperability (2020s-emerging)
• common healthcare knowledge graphs
• LLM-driven semantic mapping
• automated record linkage & ontology alignment
• uncertainty bounded reasoning
• organisational interoperability

History


1. Standards
1898 - ICD Prototype (Bertillon Classification)
First global attempt to classify diseases
1966 - CPT Introduced
Billing/procedural coding adopted in US
1975 - SNOMED (Systematised Nomenclature of Medicine)
First attempt at systemic clinical vocabulary
1990 - HL7 v2.x
First widely adopted messaging format
1992 - Arden Syntax
Early clinical logic modules
1994 - LOINC v1
Standard for lab and observation identifiers
1998 - IHE Profiles launched
Integration testing and profiles for interoperability workflows
2000 - CDA (Clinical Document Architecture)
XML based structured exchange (2003)
2003 - openEHR 1.0
Dual model architecture for semantics
2005 - HL7 v3
Model driven messaging (limited adoption)
2007- XDS.b (Cross enterprise document sharing)
IHE registry/ repository ecosystem
2014 - FHIR v1
Modern API centric interoperability
2020- AI and KG based interoperability emerges
2. Systems
1960s-1970s Early clinical data processing systems
Local, monolithic, bespoke
1970s-1980s First EMRs (Regenstrief, VA)
Digitised records but isolated. No standard data structures
1990s Client server EHR systems
Vendors like EPIC, Cerner. Still proprietary schemas, but wider adoption
2000s Web enabled EHRs + early hospital integration
Basic interoperability through HL7 v2 messaging (lab, radiology)
2010s Cloud EHRs + Data Lakes
Healthcare moves into cloud hosting, data lakes & analytics platforms
2010-2020s Multi modal systems: LIMS, imaging, genomic systems
highly heterogeneous data flows
2020s Federated systems, secure enclaves, AI pipelines
secure MPC, EHR linked knowledge graphs, semantic AI pipelines, digital health ecosystems
3. Interoperability
1st Generation - Pre Interoperability Era (1960s-1995s)
• no interoperability
• fully proprietary EHR/EMR systems
• no terminology consensus
• no exchange formats
2nd Generation - Standards-Based Interoperability (1990s-2020s)
• messaging standards (HL7 v2, v3), API (FHIR)
• terminology standards (ICD, SNOMED, LOINC)
• content standards (CDA, openEHR)
• no unified semantics
• semantics are mapped manually, inconsistently
3rd Generation - AI-driven Interoperability (2020s-emerging)
• common healthcare knowledge graphs
• LLM-driven semantic mapping
• automated record linkage & ontology alignment
• uncertainty bounded reasoning
• organisational interoperability

Informatics Models

Predate computers …

- Jean Fernel, produced the first recognised classification of diseases structured according to contemporary principles of scientific empiricism in the Universa Medicina, published in 1554
- Carolus Linnaeus (famed for his botanical taxonomy) also classified the animal and mineral kingdoms, and attempted the same for diseases

19th Century Statistics

First attempts at structured vocabularies

- 1839. First Classification of Mortalities, culminating in the International List of Causes of Death (ILCD)
- ILCD was systematically expanded by Jacques Bertillon into a classification: Bertillon Classification of Causes of Death (1880s – 1890s)
- Adopted by the WHO in 1948 to become ICD-6

First Computer Systems

1955 – 1970s: Foundations of the modern EHR

- MUMPS, otherwise known as the Massachusetts General Hospital Utility Multi-Programming System.
- Problem-Oriented Medical Records (POMR) that changed how doctors recorded patient interactions.

First Computer Systems

1970s – 1995: Foundations of the modern EHR

- EPIC Cosmos
- SNOMED-CT in 1975
- Slow development of other standards: LOINC, CPT
- Transition from National Databases (e.g. Medline) to Site-Level (provider) Systems
- Proprietary Information Models, leading to a sprawl of Information Models.

SOTA early 21st Century

SOTA early 21st Century

- Rapid acceleration in Digitalization of medical records and health systems, spurred on by large government investments

SOTA early 21st Century

- Rapid acceleration in Digitalization of medical records and health systems, spurred on by large government investments
- Vendors gain prominence, such as Epic and Cerner, with proprietary information models

SOTA early 21st Century

- Rapid acceleration in Digitalization of medical records and health systems, spurred on by large government investments
- Vendors gain prominence, such as Epic and Cerner, with proprietary information models
- Shift towards web-based systems starts

SOTA early 21st Century

- Rapid acceleration in Digitalization of medical records and health systems, spurred on by large government investments
- Vendors gain prominence, such as Epic and Cerner, with proprietary information models
- Shift towards web-based systems starts
- Interoperability becomes a concern:
    - Standards based interoperability: imposition of semantics top-down
    - More standards developed

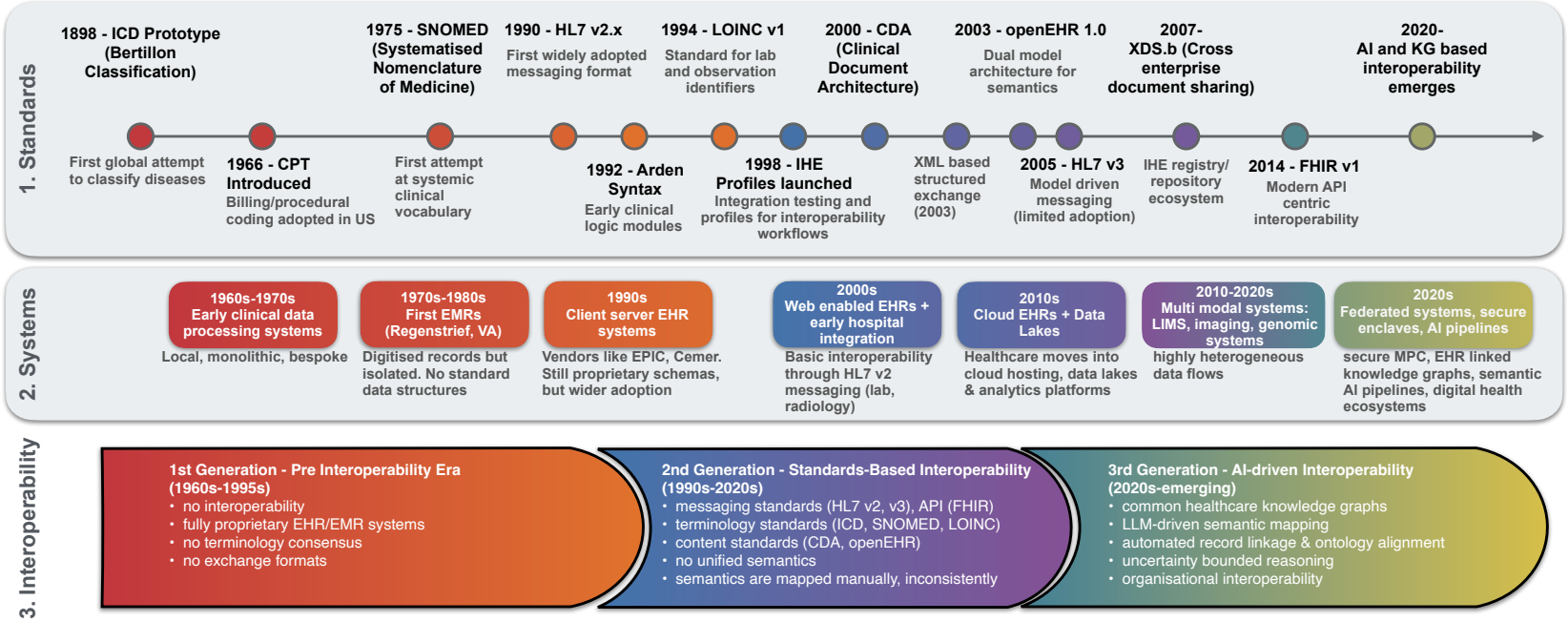
PART 02

Second Generation

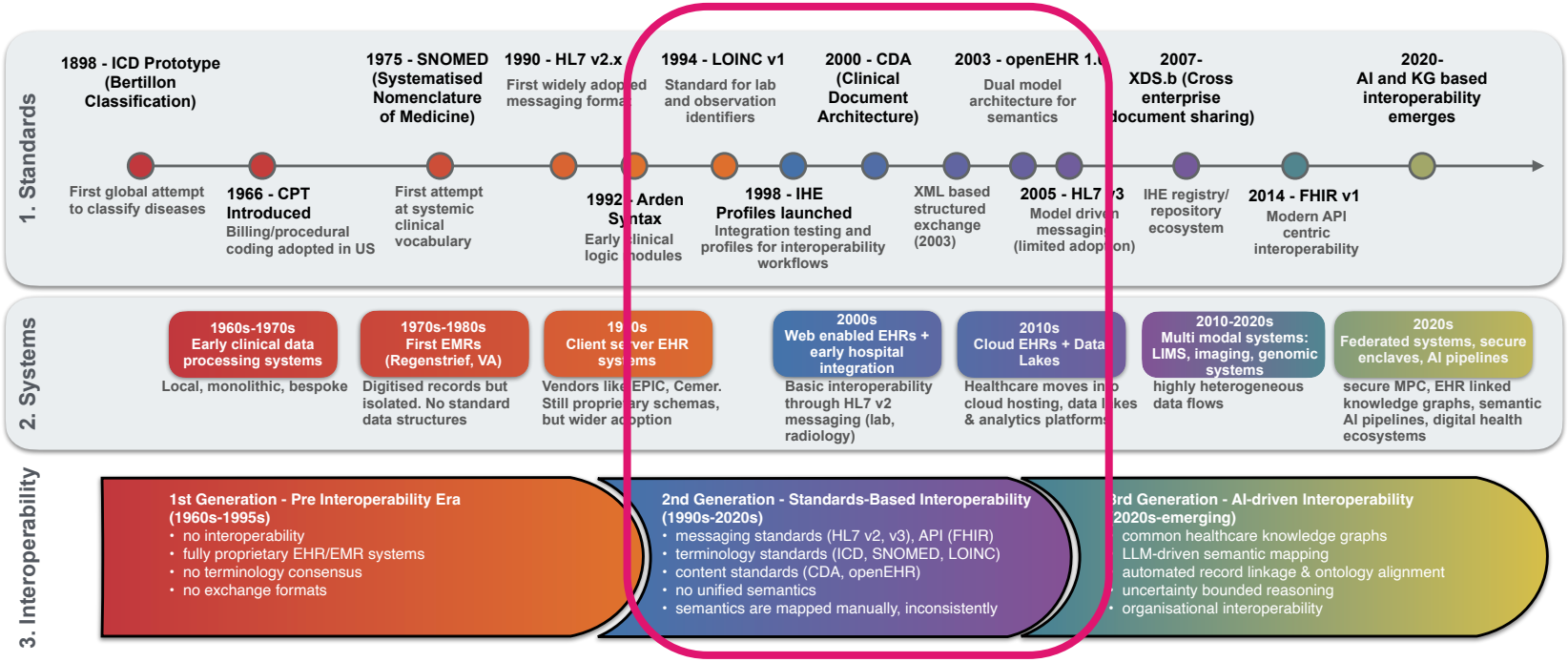
Andra Ionescu

2

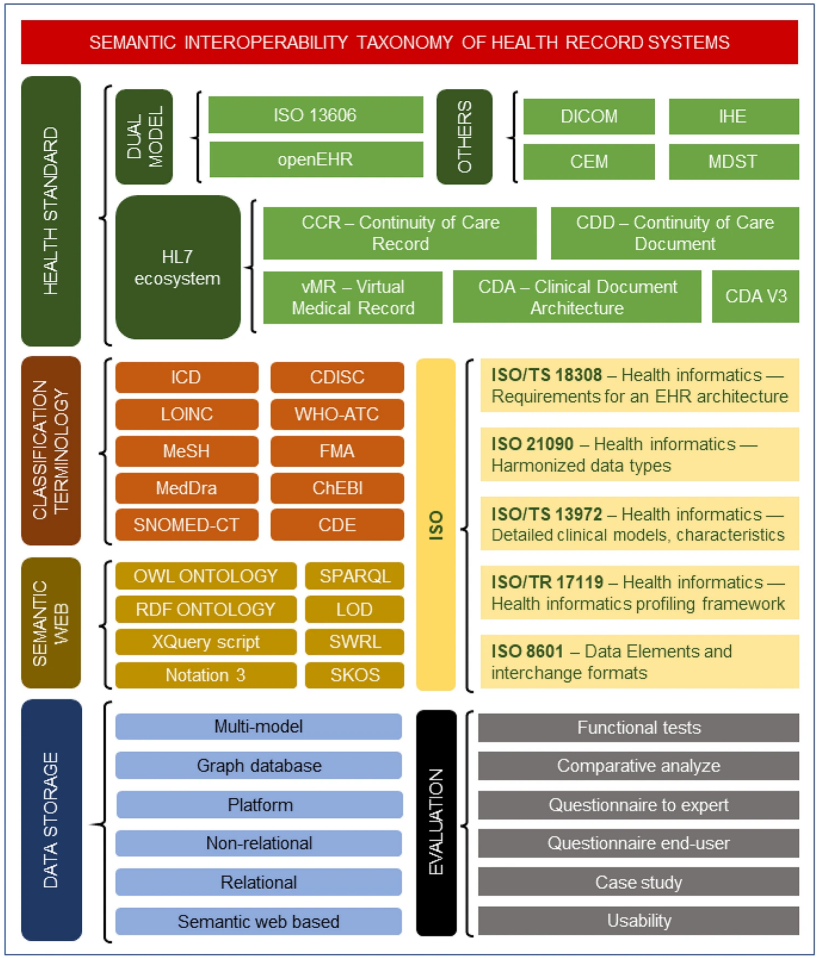
History


1. Standards
1898 - ICD Prototype (Bertillon Classification)
First global attempt to classify diseases
1966 - CPT Introduced
Billing/procedural coding adopted in US
1975 - SNOMED (Systematised Nomenclature of Medicine)
First attempt at systemic clinical vocabulary
1990 - HL7 v2.x
First widely adopted messaging format
1992 - Arden Syntax
Early clinical logic modules
1994 - LOINC v1
Standard for lab and observation identifiers
1998 - IHE Profiles launched
Integration testing and profiles for interoperability workflows
2000 - CDA (Clinical Document Architecture)
XML based structured exchange (2003)
2003 - openEHR 1.0
Dual model architecture for semantics
2005 - HL7 v3
Model driven messaging (limited adoption)
2007- XDS.b (Cross enterprise document sharing)
IHE registry/ repository ecosystem
2014 - FHIR v1
Modern API centric interoperability
2020- AI and KG based interoperability emerges
2. Systems
1960s-1970s Early clinical data processing systems
Local, monolithic, bespoke
1970s-1980s First EMRs (Regenstrief, VA)
Digitised records but isolated. No standard data structures
1990s Client server EHR systems
Vendors like EPIC, Cemer. Still proprietary schemas, but wider adoption
2000s Web enabled EHRs + early hospital integration
Basic interoperability through HL7 v2 messaging (lab, radiology)
2010s Cloud EHRs + Data Lakes
Healthcare moves into cloud hosting, data lakes & analytics platforms
2010-2020s Multi modal systems: LIMS, imaging, genomic systems
highly heterogeneous data flows
2020s Federated systems, secure enclaves, AI pipelines
secure MPC, EHR linked knowledge graphs, semantic AI pipelines, digital health ecosystems
3. Interoperability
1st Generation - Pre Interoperability Era (1960s-1995s)
• no interoperability
• fully proprietary EHR/EMR systems
• no terminology consensus
• no exchange formats
2nd Generation - Standards-Based Interoperability (1990s-2020s)
• messaging standards (HL7 v2, v3), API (FHIR)
• terminology standards (ICD, SNOMED, LOINC)
• content standards (CDA, openEHR)
• no unified semantics
• semantics are mapped manually, inconsistently
3rd Generation - AI-driven Interoperability (2020s-emerging)
• common healthcare knowledge graphs
• LLM-driven semantic mapping
• automated record linkage & ontology alignment
• uncertainty bounded reasoning
• organisational interoperability

History


1. Standards
1898 - ICD Prototype (Bertillon Classification)
First global attempt to classify diseases
1966 - CPT Introduced
Billing/procedural coding adopted in US
1975 - SNOMED (Systematised Nomenclature of Medicine)
First attempt at systemic clinical vocabulary
1990 - HL7 v2.x
First widely adopted messaging forma
1992 - Arden Syntax
Early clinical logic modules
1994 - LOINC v1
Standard for lab and observation identifiers
1998 - IHE Profiles launched
Integration testing and profiles for interoperability workflows
2000 - CDA (Clinical Document Architecture)
XML based structured exchange (2003)
2003 - openEHR 1.0
Dual model architecture for semantics
2005 - HL7 v3
Model driven messaging (limited adoption)
2007- XDS.b (Cross enterprise document sharing)
IHE registry/ repository ecosystem
2014 - FHIR v1
Modern API centric interoperability
2020- AI and KG based interoperability emerges
2. Systems
1960s-1970s Early clinical data processing systems
Local, monolithic, bespoke
1970s-1980s First EMRs (Regenstrief, VA)
Digitised records but isolated. No standard data structures
1990s Client server EHR systems
Vendors like EPIC, Cemer. Still proprietary schemas, but wider adoption
2000s Web enabled EHRs + early hospital integration
Basic interoperability through HL7 v2 messaging (lab, radiology)
2010s Cloud EHRs + Data Lakes
Healthcare moves into cloud hosting, data lakes & analytics platforms
2010-2020s Multi modal systems: LIMS, imaging, genomic systems
highly heterogeneous data flows
2020s Federated systems, secure enclaves, AI pipelines
secure MPC, EHR linked knowledge graphs, semantic AI pipelines, digital health ecosystems
3. Interoperability
1st Generation - Pre Interoperability Era (1960s-1995s)
• no interoperability
• fully proprietary EHR/EMR systems
• no terminology consensus
• no exchange formats
2nd Generation - Standards-Based Interoperability (1990s-2020s)
• messaging standards (HL7 v2, v3), API (FHIR)
• terminology standards (ICD, SNOMED, LOINC)
• content standards (CDA, openEHR)
• no unified semantics
• semantics are mapped manually, inconsistently
3rd Generation - AI-driven Interoperability (2020s-emerging)
• common healthcare knowledge graphs
• LLM-driven semantic mapping
• automated record linkage & ontology alignment
• uncertainty bounded reasoning
• organisational interoperability

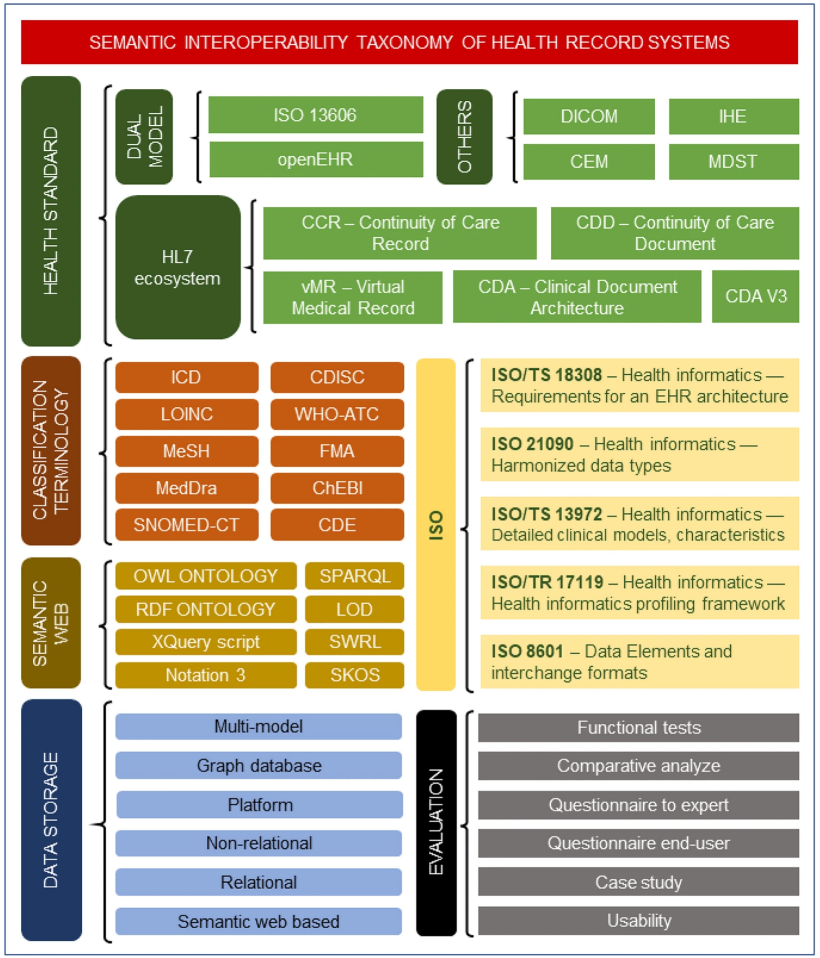
Taxonomy



de Mello, B.H., Rigo, S.J., da Costa, C.A. et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol. ***12****, 255–272 (2022).*

Taxonomy

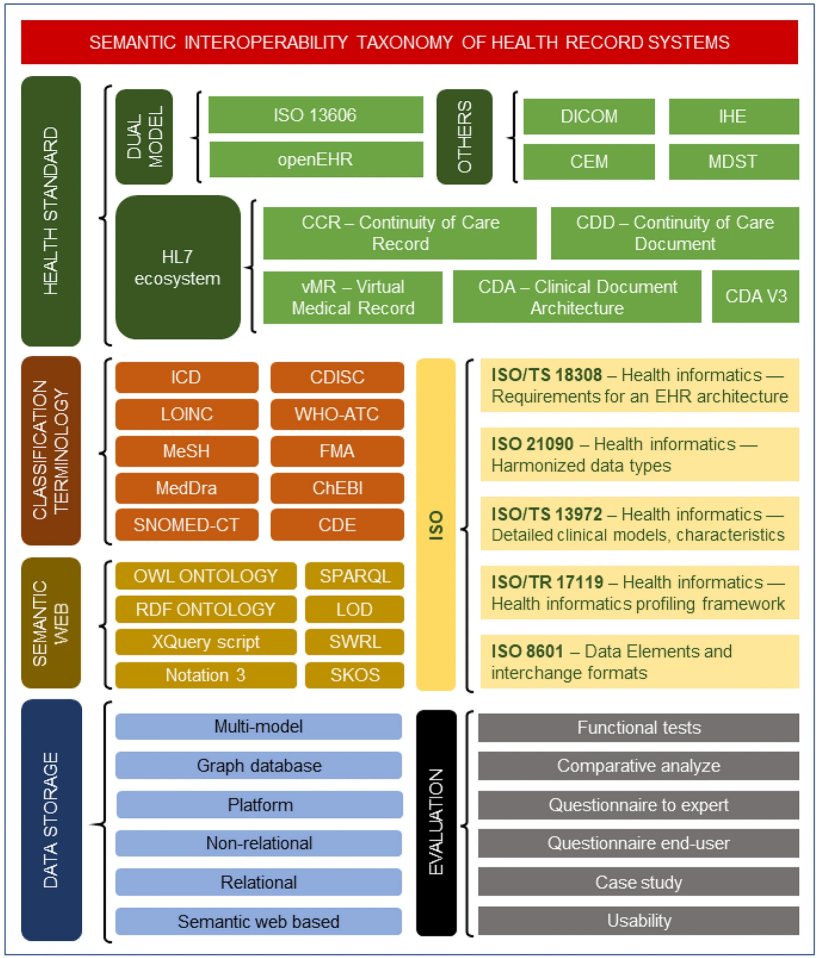
- Interoperability works by conforming to standards



de Mello, B.H., Rigo, S.J., da Costa, C.A. et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol. ***12****, 255–272 (2022).*

Taxonomy

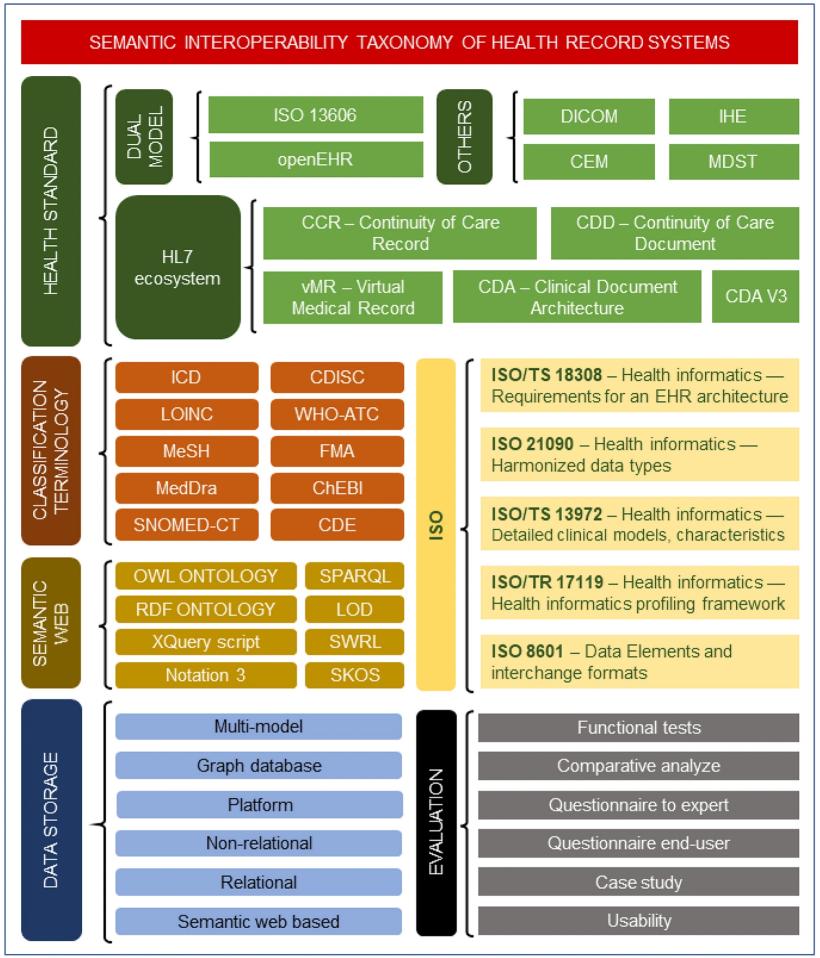
- Interoperability works by conforming to standards
- Else, *N* connectors need to be developed to communicate with the different systems



de Mello, B.H., Rigo, S.J., da Costa, C.A. et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol. ***12****, 255–272 (2022).*

Taxonomy

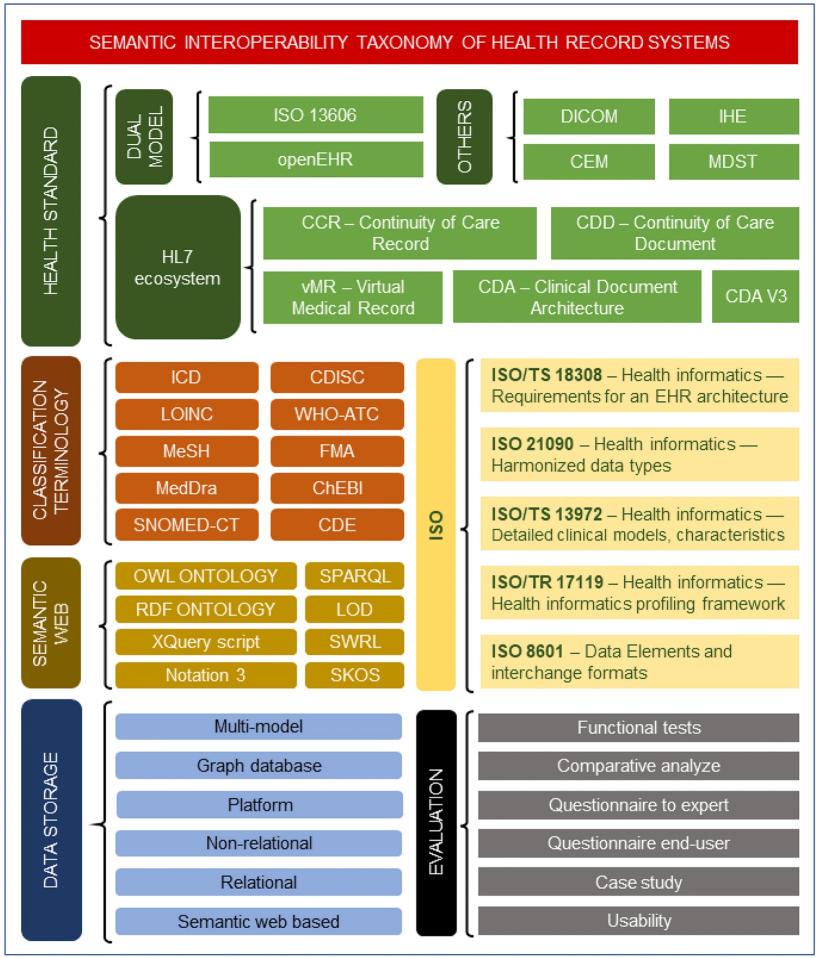
- Interoperability works by conforming to standards
- Else, *N* connectors need to be developed to communicate with the different systems
- However, there is no global consensus for standards



de Mello, B.H., Rigo, S.J., da Costa, C.A. et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol. ***12****, 255–272 (2022).*

Taxonomy

- Interoperability works by conforming to standards
- Else, *N* connectors need to be developed to communicate with the different systems
- However, there is no global consensus for standards



de Mello, B.H., Rigo, S.J., da Costa, C.A. et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol. ***12****, 255–272 (2022).*

Taxonomy

Taxonomy

Exchange Standards

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Information (domain) Models

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Common Data Models

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Common Data Models

- Subclass of Information Models

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Common Data Models

- Subclass of Information Models
- It is standardised as a shared target across institutions (aka a mediated schema)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Common Data Models

- Subclass of Information Models
- It is standardised as a shared target across institutions (aka a mediated schema)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Terminology Systems

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Common Data Models

- Subclass of Information Models
- It is standardised as a shared target across institutions (aka a mediated schema)

Terminology Systems

- Standardised vocabulary of clinical meanings

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Common Data Models

- Subclass of Information Models
- It is standardised as a shared target across institutions (aka a mediated schema)

Terminology Systems

- Standardised vocabulary of clinical meanings
- Defines the content (e.g., instance-level)

Taxonomy

Exchange Standards

- How data is serialised and transmitted between systems (aka a messaging protocol, the serialisation)

Information (domain) Models

- How clinical information is structured, a conceptual logical model (aka the schema)

Common Data Models

- Subclass of Information Models
- It is standardised as a shared target across institutions (aka a mediated schema)

Terminology Systems

- Standardised vocabulary of clinical meanings
- Defines the content (e.g., instance-level)
- Referenced ontology

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays

v2 example

```
MSH|^~\&|ADT_SENDER|HOSP_A|ADT_RECEIVER|HOSP_B|
202607291200||ADT^A01|MSG00001|P|2.5.1
EVN|A01|202607291200
PID|1||123456^^^HOSP_A^MR||Andersson^Anna^^^||
19800101|F|||Storgatan 1^^Stockholm^^11122^SE
PV1|1|I|WARD3^301^1||||||MED|||ADM
```

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays
- Limited semantic meaning, implied data model, bespoke interfaces, not scalable

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays
- Limited semantic meaning, implied data model, bespoke interfaces, not scalable

v3

Deployed 2005 — built based on a Reference Information Model (RIM) and encoded in XML

Messag ds

Health Level Seven (HL stems

v2

Deployed 1987 — the most c
standard in the world

- Pipe-and-hat delimited synta
fields, very flexible
- Legacy protocol, still used no
- Limited semantic meaning, i
model, bespoke interfaces, r

```xml
<?xml version="1.0" encoding="UTF-8"?>
<PRPA_IN402001 xmlns="urn:hl7-org:v3" ITSVersion="XML_1.0">
  <id root="2.16.840.1.113883.19.1" extension="MSG00001"/>   <!-- MSH-10 control id -->
  <creationTime value="202607291200"/>                        <!-- MSH-7 timestamp -->
  <interactionId root="2.16.840.1.113883.1.6" extension="PRPA_IN402001"/>          <!-- ≈ MSH-9 message type -->
  <processingCode code="P"/>                                  <!-- MSH-11 Production -->
  <processingModeCode code="T"/>
  <acceptAckCode code="AL"/>
  <versionCode code="V3-2011"/>                               <!-- ≈ MSH-12 (a v3 version, not 2.5.1) -->
  <receiver typeCode="RCV">
    <device classCode="DEV" determinerCode="INSTANCE">
      <id extension="ADT_RECEIVER"/>                          <!-- MSH-5 -->
      <asAgent><representedOrganization>
        <id extension="HOSP_B"/>                               <!-- MSH-6 -->
      </representedOrganization></asAgent>
    </device>
  </receiver>
  <sender typeCode="SND">
    <device classCode="DEV" determinerCode="INSTANCE">
      <id extension="ADT_SENDER"/>                            <!-- MSH-3 -->
      <asAgent><representedOrganization>
        <id extension="HOSP_A"/>                               <!-- MSH-4 -->
      </representedOrganization></asAgent>
    </device>
  </sender>
  <controlActProcess classCode="CACT" moodCode="EVN">
    <code code="A01" codeSystem="2.16.840.1.113883.1.18"/>              <!-- trigger event = admission -->
    <effectiveTime value="202607291200"/>                      <!-- EVN-2 -->
    <subject typeCode="SUBJ">
      <encounterEvent classCode="ENC" moodCode="EVN">
        <id root="2.16.840.1.113883.19.2" extension="VISIT-0001"/>
        <code code="IMP" codeSystem="2.16.840.1.113883.5.4"/>          <!-- PV1-2: Inpatient encounter -->
        <statusCode code="active"/>
        <effectiveTime value="202607291200"/>
        <subject typeCode="SBJ">
          <patient classCode="PAT">
            <id extension="123456" root="HOSP_A"/>             <!-- PID-3: MR 123456 @ HOSP_A -->
            <patientPerson classCode="PSN" determinerCode="INSTANCE">
              <name>
                <family>Andersson</family>                     <!-- PID-5 family -->
                <given>Anna</given>                            <!-- PID-5 given -->
              </name>
              <administrativeGenderCode code="F" codeSystem="2.16.840.1.113883.5.1"/>       <!-- PID-8 -->
              <birthTime value="19800101"/>                    <!-- PID-7 -->
              <addr>                                           <!-- PID-11 -->
                <streetAddressLine>Storgatan 1</streetAddressLine>
                <city>Stockholm</city>
                <postalCode>11122</postalCode>
                <country>SE</country>
              </addr>
            </patientPerson>
            <providerOrganization classCode="ORG" determinerCode="INSTANCE">
              <id extension="HOSP_A"/>                         <!-- assigning authority -->
            </providerOrganization>
          </patient>
        </subject>
        <location typeCode="LOC">
          <serviceDeliveryLocation classCode="SDLOC">
            <code code="MED" codeSystem="{hospital-service-OID}"/> <!-- PV1-10 service -->
            <location classCode="PLC" determinerCode="INSTANCE">
              <name>WARD3</name>                               <!-- point of care -->
            </location>
          </serviceDeliveryLocation>
        </location>
      </encounterEvent>
    </subject>
  </controlActProcess>
</PRPA_IN402001>
```

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays
- Limited semantic meaning, implied data model, bespoke interfaces, not scalable

v3

Deployed 2005 — built based on a Reference Information Model (RIM) and encoded in XML

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays
- Limited semantic meaning, implied data model, bespoke interfaces, not scalable

v3

Deployed 2005 — built based on a Reference Information Model (RIM) and encoded in XML

- Complex, rigid: data had to be encoded based on RIM Act-Role-Entity-Participation backbone
- Hard to implement in hospital workflows
- Not backwards compatible with v2
- Not used anymore, demanded too many resources vs ROI

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays
- Limited semantic meaning, implied data model, bespoke interfaces, not scalable

v3

Deployed 2005 — built based on a Reference Information Model (RIM) and encoded in XML

- Complex, rigid: data had to be encoded based on RIM Act-Role-Entity-Participation backbone
- Hard to implement in hospital workflows
- Not backwards compatible with v2
- Not used anymore, demanded too many resources vs ROI

FHIR

Fast Healthcare Interoperability Resources — released in 2014

- Combines the strengths of v2 and v3
- Explicit semantics: breaks data into 157 resources (e.g., Patient, Observation, Condition)
- Sends data as JSON/XML over REST APIs

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays
- Limited semantic meaning, implied data model, bespoke interfaces, not scalable

v3

Deployed 2005 — built based on a Reference Information Model (RIM) and encoded in XML

- Complex, rigid: data had to be encoded based on RIM Act-Role-Entity-Participation backbone
- Hard to implement in hospital workflows
- Not backwards compatible with v2
- Not used anymore, demanded too many resources vs ROI

FHIR

Fast Healthcare Interoperability Resources — released in 2014

- Combines the strengths of v2 and v3
- Explicit semantics: breaks data into 157 resources (e.g., Patient, Observation, Condition)
- Sends data as JSON/XML over REST APIs
- Flexibility is an issue: extra fields need data processing
- Semantic challenge: mapping data to FHIR resources

M

Hea

De
sta

-
-


```
{
  "resourceType": "Bundle",
  "type": "message",
  "timestamp": "2026-07-29T12:00:00+02:00",
  "identifier": { "value": "MSG00001" },
  "entry": [
    {
      "fullUrl": "urn:uuid:header-1",
      "resource": {
        "resourceType": "MessageHeader",
        "eventCoding": {
          "system": "http://terminology.hl7.org/CodeSystem/v2-0003",
          "code": "A01",
          "display": "ADT/ACK - Admit/visit notification"
        },
        "source": {
          "name": "ADT_SENDER",
          "endpoint": "urn:hospital:HOSP_A"
        },
        "destination": [
          { "name": "ADT_RECEIVER", "endpoint": "urn:hospital:HOSP_B" }
        ],
        "focus": [ { "reference": "urn:uuid:encounter-1" } ]
      }
    },
    {
      "fullUrl": "urn:uuid:patient-1",
      "resource": {
        "resourceType": "Patient",
        "identifier": [
          {
            "type": {
              "coding": [ {
                "system": "http://terminology.hl7.org/CodeSystem/v2-0203",
                "code": "MR"
              } ]
            },
            "system": "urn:hospital:HOSP_A",
            "value": "123456"
          }
        ],
        "name": [ { "family": "Andersson", "given": [ "Anna" ] } ],
        "gender": "female",
        "birthDate": "1980-01-01",
        "address": [
          {
            "line": [ "Storgatan 1" ],
            "city": "Stockholm",
            "postalCode": "11122",
            "country": "SE"
          }
        ]
      }
    },
```



```
    {
      "fullUrl": "urn:uuid:encounter-1",
      "resource": {
        "resourceType": "Encounter",
        "status": "in-progress",
        "class": {
          "system": "http://terminology.hl7.org/CodeSystem/v3-ActCode",
          "code": "IMP",
          "display": "inpatient encounter"
        },
        "subject": { "reference": "urn:uuid:patient-1" },
        "period": { "start": "2026-07-29T12:00:00+02:00" },
        "serviceType": {
          "coding": [ {
            "system": "http://terminology.hl7.org/CodeSystem/service-type",
            "code": "MED",
            "display": "Medical service"
          } ]
        },
        "location": [
          {
            "location": { "reference": "urn:uuid:location-1" },
            "status": "active"
          }
        ],
        "serviceProvider": { "reference": "urn:uuid:org-1" }
      }
    },
    {
      "fullUrl": "urn:uuid:location-bed",
      "resource": {
        "resourceType": "Location",
        "name": "Bed 1",
        "physicalType": {
          "coding": [ {
            "system": "http://terminology.hl7.org/CodeSystem/location-physical-type",
            "code": "bd", "display": "Bed"
          } ]
        },
        "partOf": { "reference": "urn:uuid:location-ward" }
      }
    },
    {
      "fullUrl": "urn:uuid:location-ward",
      "resource": {
        "resourceType": "Location",
        "name": "WARD3",
        "physicalType": {
          "coding": [ {
            "system": "http://terminology.hl7.org/CodeSystem/location-physical-type",
            "code": "wa", "display": "Ward"
          } ]
        }
      }
    }
    {
      "fullUrl": "urn:uuid:org-1",
      "resource": {
        "resourceType": "Organization",
        "identifier": [ { "value": "HOSP_A" } ],
        "name": "Hospital A"
      }
    }
  ]
}
```

Message Exchange Standards

Health Level Seven (HL7) – exchanging clinical and administrative data between healthcare systems

v2

Deployed 1987 — the most deployed standard in the world

- Pipe-and-hat delimited syntax, lots of optional fields, very flexible
- Legacy protocol, still used nowadays
- Limited semantic meaning, implied data model, bespoke interfaces, not scalable

v3

Deployed 2005 — built based on a Reference Information Model (RIM) and encoded in XML

- Complex, rigid: data had to be encoded based on RIM Act-Role-Entity-Participation backbone
- Hard to implement in hospital workflows
- Not backwards compatible with v2
- Not used anymore, demanded too many resources vs ROI

FHIR

Fast Healthcare Interoperability Resources — released in 2014

- Combines the strengths of v2 and v3
- Explicit semantics: breaks data into 157 resources (e.g., Patient, Observation, Condition)
- Sends data as JSON/XML over REST APIs
- Flexibility is an issue: extra fields need data processing
- Semantic challenge: mapping data to FHIR resources

Message Exchange Standards

WHY?

Message Exchange Standards

WHY?

- Two systems that are built completely independent from each other agree on a data representation that can be sent and received

Message Exchange Standards

WHY?

- Two systems that are built completely independent from each other agree on a data representation that can be sent and received
- They don’t have to know the internal schema, the message just needs to adhere to the standardised format

Message Exchange Standards

WHY?

- Two systems that are built completely independent from each other agree on a data representation that can be sent and received
- They don’t have to know the internal schema, the message just needs to adhere to the standardised format

Mapping problem: mappings are manually defined

Message Exchange Standards

WHY?

- Two systems that are built completely independent from each other agree on a data representation that can be sent and received
- They don’t have to know the internal schema, the message just needs to adhere to the standardised format

- Mapping problem: mappings are manually defined
- Data loss: not all data can be sent

Message Exchange Standards

WHY?

- Two systems that are built completely independent from each other agree on a data representation that can be sent and received
- They don’t have to know the internal schema, the message just needs to adhere to the standardised format

- Mapping problem: mappings are manually defined
- Data loss: not all data can be sent
- Semantic problem: the sender doesn’t know if the receiver understands the data

Information Models

openEHR

- An architecture to store, version and query EHR
- Captures the epistemic role of data, it is the closest model to clinical practice, similar to how clinicians think
- It is based on a very strict Reference Model that contains *data types* (SECTION, ENTRY, CLUSTER, ELEMENT, etc)

Information Models

openEHR

- An architecture to store, version and query EHR
- Captures the epistemic role of data, it is the closest model to clinical practice, similar to how clinicians think
- It is based on a very strict Reference Model that contains *data types* (SECTION, ENTRY, CLUSTER, ELEMENT, etc)

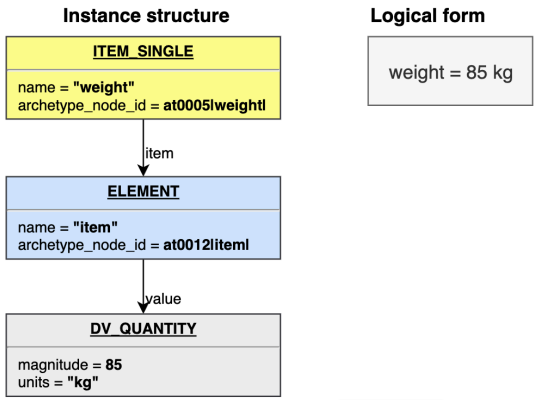


Figure 2. Instance Structure of `ITEM_SINGLE`

https://specifications.openehr.org/releases/RM/Release-1.1.0/data_structures.html

Information Models

openEHR

- An architecture to store, version and query EHR
- Captures the epistemic role of data, it is the closest model to clinical practice, similar to how clinicians think
- It is based on a very strict Reference Model that contains *data types* (SECTION, ENTRY, CLUSTER, ELEMENT, etc)

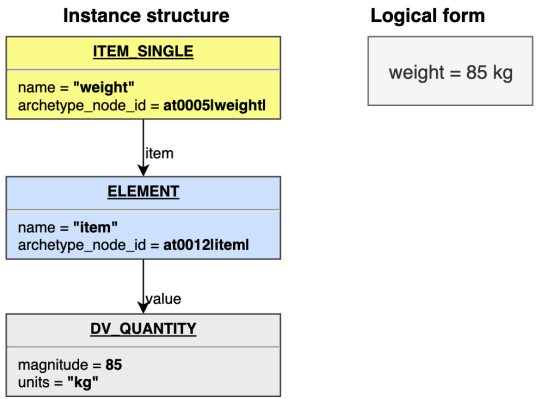


Figure 2. Instance Structure of `ITEM_SINGLE`

High complexity: modelling data before storing

https://specifications.openehr.org/releases/RM/Release-1.1.0/data_structures.html

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

openEHR

```
OBSERVATION: Blood pressure  [openEHR-EHR-OBSERVATION.blood_pressure.v2]
└─ data: HISTORY
   └─ events[0]: POINT_EVENT "Any event" @ 2026-07-29T12:00:00
      ├─ data (ITEM_TREE)                ← the measured values
      │  ├─ ELEMENT at0004 "Systolic"   = 120 mm[Hg]  (DV_QUANTITY)
      │  └─ ELEMENT at0005 "Diastolic"  =  80 mm[Hg]  (DV_QUANTITY)
      └─ state (ITEM_TREE)               ← condition at time of measurement
         └─ ELEMENT at0008 "Position"  = Sitting  (DV_CODED_TEXT)
└─ protocol (ITEM_TREE)                  ← how it was measured
   ├─ ELEMENT at1007 "Cuff size"          = Adult
   └─ ELEMENT at1008 "Location"           = Right arm  (DV_CODED_TEXT)
```

Information Models

OMOP

Observational Medical Outcomes Partnership

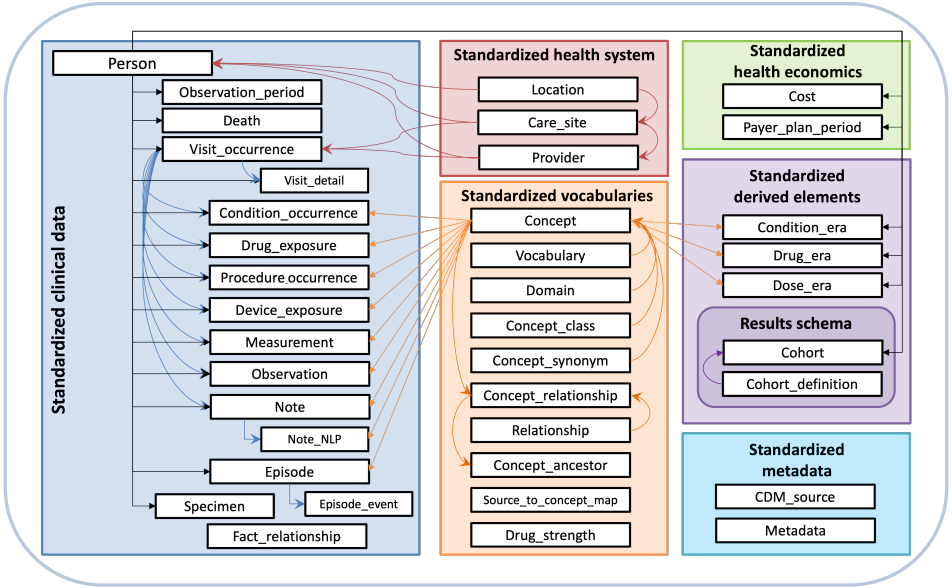
- Relational schema with standardised tables
- Clinical facts mapped to a *concept_id* from OMOP's Standardised Vocabularies
- Built for longitudinal, population-scale analysis

Information Models

OMOP

Observational Medical Outcomes Partnership

- Relational schema with standardised tables
- Clinical facts mapped to a *concept_id* from OMOP's Standardised Vocabularies
- Built for longitudinal, population-scale analysis



https://ohdsi.github.io/CommonDataModel/

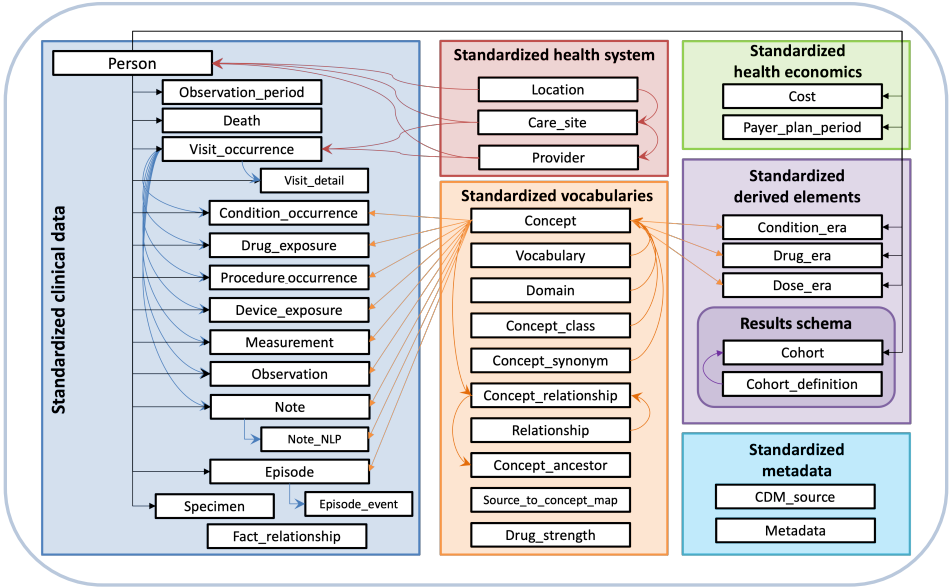
Information Models

OMOP

Observational Medical Outcomes Partnership

- Relational schema with standardised tables
- Clinical facts mapped to a *concept_id* from OMOP's Standardised Vocabularies
- Built for longitudinal, population-scale analysis

Not an exchange nor a persistence standard



https://ohdsi.github.io/CommonDataModel/

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

OMOP

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

OMOP

MEASUREMENT table
23 columns

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

OMOP

MEASUREMENT table
23 columns

measurement_id,person_id,measurement_concept_id,measurement_date,measurement_datetime,measurement_time,measurement_type_concept_id,operator_concept_id,value_as_number,value_as_concept_id,unit_concept_id,range_low,range_high,provider_id,visit_occurrence_id,visit_detail_id,measurement_source_value,measurement_source_concept_id,unit_source_value,unit_source_concept_id,value_source_value,measurement_event_id,meas_event_field_concept_id
1001,42,3004249,2026-07-29,2026-07-29 12:00:00,12:00:00,32817,,120,,8876,90,120,,5001,,8480-6,3004249,mmHg,8876,120,,
1002,42,3012888,2026-07-29,2026-07-29 12:00:00,12:00:00,32817,,80,,8876,60,80,,5001,,8462-4,3012888,mmHg,8876,80,,

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

OMOP

MEASUREMENT table
23 columns

measurement_id,person_id,measurement_concept_id,measurement_date,measurement_datetime,measurement_time,measurement_type_concept_id,operator_concept_id,value_as_number,value_as_concept_id,unit_concept_id,range_low,range_high,provider_id,visit_occurrence_id,visit_detail_id,measurement_source_value,measurement_source_concept_id,unit_source_value,unit_source_concept_id,value_source_value,measurement_event_id,meas_event_field_concept_id

1001,42,3004249,2026-07-29,2026-07-29 12:00:00,12:00:00,32817,,120,,8876,90,120,,5001,,8480-6,3004249,mmHg,8876,120,,

1002,42,3012888,2026-07-29,2026-07-29 12:00:00,12:00:00,32817,,80,,8876,60,80,,5001,,8462-4,3012888,mmHg,8876,80,,

Information Models

Information Models

i2b2

Information Models

i2b2

Information Models

i2b2

Information Models

i2b2

Informatics for Integrating Biology and the Bedside

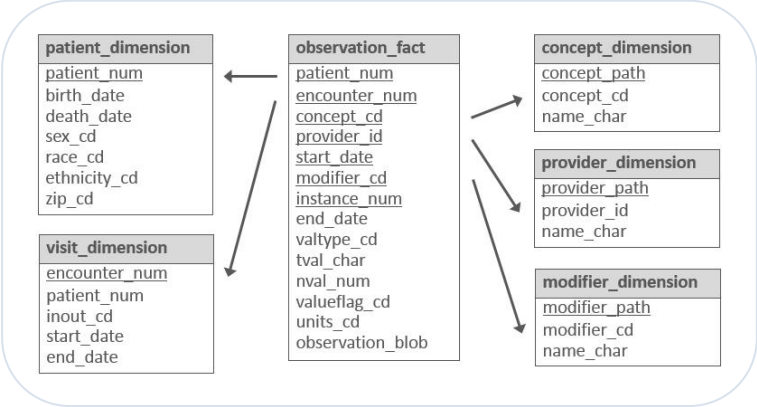
- Common data model for data warehouse development
- Star schema with one large fact table of atomic observations with dimension tables

Information Models

i2b2

Informatics for Integrating Biology and the Bedside

- Common data model for data warehouse development
- Star schema with one large fact table of atomic observations with dimension tables



i2b2 star schema — fact table with dimension tables (https://i2b2.atlassian.net/wiki/spaces/BUN/pages/207290415/)

https://freepngimg.com/png/104629-picture-r2-d2-free-transparent-image-hd/icon

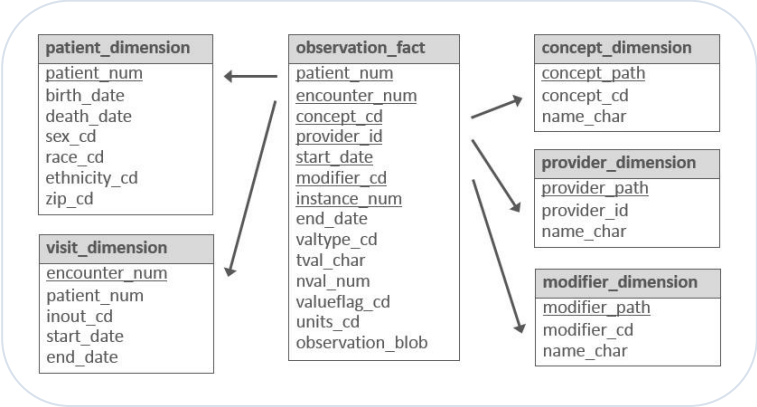
Information Models

i2b2

Informatics for Integrating Biology and the Bedside

- Common data model for data warehouse development
- Star schema with one large fact table of atomic observations with dimension tables

- Narrow analytic scope: cohort discovery



i2b2 star schema — fact table with dimension tables (https://i2b2.atlassian.net/wiki/spaces/BUN/pages/207290415/)

https://freepngimg.com/png/104629-picture-r2-d2-free-transparent-image-hd/icon

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

i2b2

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

i2b2

OBSERVATION_FACT table
22 columns

Blood Pressure Example

Example: 120/80 mmHg, taken 2026-07-29, patient id 42, during visit 5001

i2b2

OBSERVATION_FACT table
22 columns

```
encounter_num,patient_num,concept_cd,provider_id,start_date,modifier_cd,instance_num,valtype_cd,tval_char,nval_num,valueflag_cd,quantity_num,units_cd,end_date,location_cd,confidence_num,observation_blob,update_date,download_date,import_date,sourcesystem_cd,upload_id
5001,42,LOINC:8480-6,@,2026-07-29 12:00:00,@,1,N,E,120,,,mmHg,2026-07-29 12:00:00,,,,2026-07-29,2026-07-29,2026-07-29,EHR,1
5001,42,LOINC:8462-4,@,2026-07-29 12:00:00,@,1,N,E,80,,,mmHg,2026-07-29 12:00:00,,,,2026-07-29,2026-07-29,2026-07-29,EHR,1
```

Information Models

WHY?

Information Models

WHY?

- The information models are diverse, no one size fits all

Information Models

WHY?

- The information models are diverse, no one size fits all
- The information models still need mapping to the exchange protocols

Information Models

WHY?

- The information models are diverse, no one size fits all
- The information models still need mapping to the exchange protocols

- They don’t eliminate mapping, they reduce the load: map once to the standard and then all systems talk with each other

Information Models

WHY?

- The information models are diverse, no one size fits all
- The information models still need mapping to the exchange protocols

- They don’t eliminate mapping, they reduce the load: map once to the standard and then all systems talk with each other
- They predetermine the syntax, structure, semantics – they decide what to NOT map

Information Models

WHY?

- The information models are diverse, no one size fits all
- The information models still need mapping to the exchange protocols

- They don’t eliminate mapping, they reduce the load: map once to the standard and then all systems talk with each other
- They predetermine the syntax, structure, semantics – they decide what to NOT map
- They bind to shared terminology systems

Data Integration Challenges

Data Integration Challenges

THE CHALLENGE

Absence of a stable mediated schema

Data Integration Challenges

THE CHALLENGE

Absence of a stable mediated schema

THE SOLUTION

Common Data Model (CDM) is the attempt to impose a mediated schema (OMOP, PCORnet, Sentinel, etc.)

Lenzerini. 2002 "Data Integration: A Theoretical Perspective”
Doan et al. 2012 “Principles of Data Integration”
Garza et al. 2019 “Evaluating common data models for use with a longitudinal community registry”

Data Integration Challenges

THE CHALLENGE

Absence of a stable mediated schema

THE SOLUTION

Common Data Model (CDM) is the attempt to impose a mediated schema (OMOP, PCORnet, Sentinel, etc.)

YES, BUT …

Lenzerini. 2002 "Data Integration: A Theoretical Perspective"
Doan et al. 2012 "Principles of Data Integration"
Garza et al. 2019 "Evaluating common data models for use with a longitudinal community registry"

Data Integration Challenges

THE CHALLENGE

Absence of a stable mediated schema

THE SOLUTION

Common Data Model (CDM) is the attempt to impose a mediated schema (OMOP, PCORnet, Sentinel, etc.)

YES, BUT …

1. No CDM is complete (Garza et al.)

Lenzerini. 2002 "Data Integration: A Theoretical Perspective"
Doan et al. 2012 "Principles of Data Integration"
Garza et al. 2019 "Evaluating common data models for use with a longitudinal community registry"

Data Integration Challenges

THE CHALLENGE

Absence of a stable mediated schema

THE SOLUTION

Common Data Model (CDM) is the attempt to impose a mediated schema (OMOP, PCORnet, Sentinel, etc.)

YES, BUT …

- No CDM is complete (Garza et al.)
- Concepts and vocabularies change continuously

Lenzerini. 2002 "Data Integration: A Theoretical Perspective"
Doan et al. 2012 "Principles of Data Integration"
Garza et al. 2019 "Evaluating common data models for use with a longitudinal community registry"

Data Integration Challenges

THE CHALLENGE

Absence of a stable mediated schema

THE SOLUTION

Common Data Model (CDM) is the attempt to impose a mediated schema (OMOP, PCORnet, Sentinel, etc.)

YES, BUT …

- No CDM is complete (Garza et al.)
- Concepts and vocabularies change continuously
- Semantic loss due to incomplete or incorrect mappings

Lenzerini. 2002 "Data Integration: A Theoretical Perspective"
Doan et al. 2012 "Principles of Data Integration"
Garza et al. 2019 "Evaluating common data models for use with a longitudinal community registry"

Data Integration Challenges

Data Integration Challenges

THE CHALLENGE

Incomplete, institution-specific, poorly maintained mappings

Data Integration Challenges

THE CHALLENGE

Incomplete, institution-specific, poorly maintained mappings

THE DM SOLUTION

Schema Matching (**COMA, Cupid, Similarity Flooding, Valentine**, etc.) + Entity Resolution (**DeepMatcher, DeepER, Ditto**, etc.)

Do & Rahm. 2002 (COMA), Madhavan et al. 2001 (Cupid), Melnik et al. 2002 (Similarity Flooding), Koutras et al. 2021 (Valentine)
Mudgal et al. 2018 (DeepMatcher), Ebraheem et al. 2018 (DeepER), Li at al. 2020 (Ditto)
Rahm & Bernstein. 2001 "A survey of approaches to automatic schema matching"

Data Integration Challenges

THE CHALLENGE

Incomplete, institution-specific, poorly maintained mappings

THE DM SOLUTION

Schema Matching (**COMA, Cupid, Similarity Flooding, Valentine**, etc.) + Entity Resolution (**DeepMatcher, DeepER, Ditto**, etc.)

YES, BUT …

Do & Rahm. 2002 (COMA), Madhavan et al. 2001 (Cupid), Melnik et al. 2002 (Similarity Flooding), Koutras et al. 2021 (Valentine)
Mudgal et al. 2018 (DeepMatcher), Ebraheem et al. 2018 (DeepER), Li at al. 2020 (Ditto)
Rahm & Bernstein. 2001 "A survey of approaches to automatic schema matching"

Data Integration Challenges

THE CHALLENGE

Incomplete, institution-specific, poorly maintained mappings

THE DM SOLUTION

Schema Matching (**COMA, Cupid, Similarity Flooding, Valentine**, etc.) + Entity Resolution (**DeepMatcher, DeepER, Ditto**, etc.)

YES, BUT …

1. Automated matchers reduce the manual load, but don’t eliminate it

Do & Rahm. 2002 (COMA), Madhavan et al. 2001 (Cupid), Melnik et al. 2002 (Similarity Flooding), Koutras et al. 2021 (Valentine)
Mudgal et al. 2018 (DeepMatcher), Ebraheem et al. 2018 (DeepER), Li at al. 2020 (Ditto)
Rahm & Bernstein. 2001 "A survey of approaches to automatic schema matching"

Data Integration Challenges

THE CHALLENGE

Incomplete, institution-specific, poorly maintained mappings

THE DM SOLUTION

Schema Matching (**COMA, Cupid, Similarity Flooding, Valentine**, etc.) + Entity Resolution (**DeepMatcher, DeepER, Ditto**, etc.)

YES, BUT …

- Automated matchers reduce the manual load, but don't eliminate it
- Schema and instances are not enough, semantics matter

Do & Rahm. 2002 (COMA), Madhavan et al. 2001 (Cupid), Melnik et al. 2002 (Similarity Flooding), Koutras et al. 2021 (Valentine)
Mudgal et al. 2018 (DeepMatcher), Ebraheem et al. 2018 (DeepER), Li at al. 2020 (Ditto)
Rahm & Bernstein. 2001 "A survey of approaches to automatic schema matching"

Data Integration Challenges

Data Integration Challenges

THE CHALLENGE

Schema heterogeneity across EHRs — difficulty finding related tables

Data Integration Challenges

THE CHALLENGE

Schema heterogeneity across EHRs — difficulty finding related tables

THE DM SOLUTION

Dataset discovery methods for finding unionable and joinable tables (**Aurum, D3L, Starmie, JOSIE, TUS, PEXESO, DeepJoin, LSH Ensemble, Seeping Semantics**, etc.)

Fernandez et al. 2018 (Aurum), Bogatu et al. 2020 (D3L), Fang et al. 2023 (Starmie), Zhu et al. 2019 (JOSIE), Nargesian et al. 2018 (TUS), Dong et al. 2021 (PEXESO), Dong at al. 2023 (DeepJoin), Zhu et al. 2016 (LSH Ensamble), Fernandez et al. 2018 (Seeping Semantics), Zhang et al. 2021 (SMAT)
Abedjan et al. 2025 “Data Discovery in Data Lakes: Operations, Indexes, Systems”

Data Integration Challenges

THE CHALLENGE

Schema heterogeneity across EHRs — difficulty finding related tables

THE DM SOLUTION

Dataset discovery methods for finding unionable and joinable tables (**Aurum, D3L, Starmie, JOSIE, TUS, PEXESO, DeepJoin, LSH Ensemble, Seeping Semantics**, etc.)

YES, BUT …

Fernandez et al. 2018 (Aurum), Bogatu et al. 2020 (D3L), Fang et al. 2023 (Starmie), Zhu et al. 2019 (JOSIE), Nargesian et al. 2018 (TUS), Dong et al. 2021 (PEXESO), Dong at al. 2023 (DeepJoin), Zhu et al. 2016 (LSH Ensamble), Fernandez et al. 2018 (Seeping Semantics), Zhang et al. 2021 (SMAT)
Abedjan et al. 2025 “Data Discovery in Data Lakes: Operations, Indexes, Systems”

Data Integration Challenges

THE CHALLENGE

Schema heterogeneity across EHRs — difficulty finding related tables

THE DM SOLUTION

Dataset discovery methods for finding unionable and joinable tables (**Aurum, D3L, Starmie, JOSIE, TUS, PEXESO, DeepJoin, LSH Ensemble, Seeping Semantics**, etc.)

YES, BUT …

1. Data access, federated learning: raw patient instances cannot cross institutional boundaries (Abedjan et al.)

Fernandez et al. 2018 (Aurum), Bogatu et al. 2020 (D3L), Fang et al. 2023 (Starmie), Zhu et al. 2019 (JOSIE), Nargesian et al. 2018 (TUS), Dong et al. 2021 (PEXESO), Dong at al. 2023 (DeepJoin), Zhu et al. 2016 (LSH Ensamble), Fernandez et al. 2018 (Seeping Semantics), Zhang et al. 2021 (SMAT)
Abedjan et al. 2025 “Data Discovery in Data Lakes: Operations, Indexes, Systems”

Data Integration Challenges

THE CHALLENGE

Schema heterogeneity across EHRs — difficulty finding related tables

THE DM SOLUTION

Dataset discovery methods for finding unionable and joinable tables (**Aurum, D3L, Starmie, JOSIE, TUS, PEXESO, DeepJoin, LSH Ensemble, Seeping Semantics**, etc.)

YES, BUT …

- Data access, federated learning: raw patient instances cannot cross institutional boundaries (Abedjan et al.)
- Privacy preserving discovery: schema level is preferred to prevent data leakage (SMAT)

Fernandez et al. 2018 (Aurum), Bogatu et al. 2020 (D3L), Fang et al. 2023 (Starmie), Zhu et al. 2019 (JOSIE), Nargesian et al. 2018 (TUS), Dong et al. 2021 (PEXESO), Dong at al. 2023 (DeepJoin), Zhu et al. 2016 (LSH Ensamble), Fernandez et al. 2018 (Seeping Semantics), Zhang et al. 2021 (SMAT)
Abedjan et al. 2025 “Data Discovery in Data Lakes: Operations, Indexes, Systems”

Terminology Systems

Terminology Systems

Vocabularies

Terminology Systems

Vocabularies

Terminologies

Terminology Systems

Vocabularies

Terminologies

Classifications

Terminology Systems

Vocabularies

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

RxNorm

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

RxNorm

q: paracetamol

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

RxNorm

q: paracetamol *56* results

Terminology Systems

Vocabularies

— finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

RxNorm

q: paracetamol

56 results

H Rx S	**acetaminophen (56)**
H Rx S	*8 HR acetaminophen 650 MG Extended Release Oral Tablet*
H Rx S	*50 ML acetaminophen 10 MG/ML Injection*
H Rx S	*65 ML acetaminophen 10 MG/ML Injection*
H Rx S	*100 ML acetaminophen 10 MG/ML Injection*
S	*acetaminophen 16 MG/ML Oral Solution*
H Rx S	*acetaminophen 16.7 MG/ML Oral Solution*
S	*acetaminophen 20 MG/ML Oral Suspension*
S	*acetaminophen 24 MG/ML Oral Solution*
S	*acetaminophen 24 MG/ML Oral Suspension*
S	*acetaminophen 25 MG/ML Oral Suspension*
H Rx S	*acetaminophen 32 MG/ML Oral Solution*
H Rx S	*acetaminophen 32 MG/ML Oral Suspension*
H Rx S	*acetaminophen 33.3 MG/ML Oral Solution*
H S	*acetaminophen 45 MG/ML Oral Solution*
S	*acetaminophen 48.2 MG/ML Oral Solution*
S	*acetaminophen 50 MG/ML Oral Solution*

Terminology Systems

Vocabularies

— finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

RxNorm

q: paracetamol *56* results

H Rx S	**acetaminophen (56)**
H Rx S	*8 HR acetaminophen 650 MG Extended Release Oral Tablet*
H Rx S	*50 ML acetaminophen 10 MG/ML Injection*
H Rx S	*65 ML acetaminophen 10 MG/ML Injection*
H Rx S	*100 ML acetaminophen 10 MG/ML Injection*
S	*acetaminophen 16 MG/ML Oral Solution*
H Rx S	*acetaminophen 16.7 MG/ML Oral Solution*
S	*acetaminophen 20 MG/ML Oral Suspension*
S	*acetaminophen 24 MG/ML Oral Solution*
S	*acetaminophen 24 MG/ML Oral Suspension*
S	*acetaminophen 25 MG/ML Oral Suspension*
H Rx S	*acetaminophen 32 MG/ML Oral Solution*
H Rx S	*acetaminophen 32 MG/ML Oral Suspension*
H Rx S	*acetaminophen 33.3 MG/ML Oral Solution*
H S	*acetaminophen 45 MG/ML Oral Solution*
S	*acetaminophen 48.2 MG/ML Oral Solution*
S	*acetaminophen 50 MG/ML Oral Solution*

H Rx S	**Acephen (3)**
H Rx S	*Acephen 120 MG Rectal Suppository*
H Rx S	*Acephen 325 MG Rectal Suppository*
H Rx S	*Acephen 650 MG Rectal Suppository*
H Rx M	**Acetadryl (1)**
H Rx M	*Acetadryl 500 MG / 25 MG Oral Tablet*
H Rx M	**Allzital (1)**
H Rx M	*Allzital 25 MG / 325 MG Oral Tablet*
H Rx M	**Anabar (1)**
H Rx M	*Anabar 300 MG / 20 MG / 200 MG Oral Tablet*
H Rx M	**Anacin Advanced Headache Formula (1)**
H Rx M	*ANACIN Advanced Headache Formula 250 MG*
H Rx S	**Anacin AF (1)**
H Rx S	*ANACIN Aspirin Free 500 MG Oral Tablet*

Terminology Systems

Vocabularies

— finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

RxNorm

q: paracetamol *56* results

Characteristic	Value
RxCUI	161
Concept Name	acetaminophen
Term Type	IN

H Rx S **acetaminophen (56)**
H Rx S *8 HR acetaminophen 650 MG Extended Release Oral Tablet*
H Rx S *50 ML acetaminophen 10 MG/ML Injection*
H Rx S *65 ML acetaminophen 10 MG/ML Injection*
H Rx S *100 ML acetaminophen 10 MG/ML Injection*
S *acetaminophen 16 MG/ML Oral Solution*
H Rx S *acetaminophen 16.7 MG/ML Oral Solution*
S *acetaminophen 20 MG/ML Oral Suspension*
S *acetaminophen 24 MG/ML Oral Solution*
S *acetaminophen 24 MG/ML Oral Suspension*
S *acetaminophen 25 MG/ML Oral Suspension*
H Rx S *acetaminophen 32 MG/ML Oral Solution*
H Rx S *acetaminophen 32 MG/ML Oral Suspension*
H Rx S *acetaminophen 33.3 MG/ML Oral Solution*
H S *acetaminophen 45 MG/ML Oral Solution*
S *acetaminophen 48.2 MG/ML Oral Solution*
S *acetaminophen 50 MG/ML Oral Solution*

H Rx S **Acephen (3)**
H Rx S *Acephen 120 MG Rectal Suppository*
H Rx S *Acephen 325 MG Rectal Suppository*
H Rx S *Acephen 650 MG Rectal Suppository*
H Rx M **Acetadryl (1)**
H Rx M *Acetadryl 500 MG / 25 MG Oral Tablet*
H Rx M **Allzital (1)**
H Rx M *Allzital 25 MG / 325 MG Oral Tablet*
H Rx M **Anabar (1)**
H Rx M *Anabar 300 MG / 20 MG / 200 MG Oral Tablet*
H Rx M **Anacin Advanced Headache Formula (1)**
H Rx M *ANACIN Advanced Headache Formula 250 MG*
H Rx S **Anacin AF (1)**
H Rx S *ANACIN Aspirin Free 500 MG Oral Tablet*

https://mor.nlm.nih.gov/RxNav/search?searchBy=RXCUI&searchTerm=161

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

ATC

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

ATC

q: paracetamol

Terminology Systems

Vocabularies — finite, curated set of terms with fixed meanings used within a domain: **RxNorm** (names for clinical drugs, mostly USA), **ATC** (worldwide)

ATC

q: paracetamol

N02BE **Anilides**
Propacetamol, a prodrug of paracetamol is classified at a separate ATC 5th level in this group.
Benorilate, which is an ester of acetylsalicylic acid and paracetamol, is classified in N02BA.
Combinations of paracetamol and e.g. ibuprofen are classified in N02BE51.
Paracetamol in combination with orphenadrine (citrate) is classified in M03BC.

ATC code	Name	DDD	U	Adm.R	Note
N02BE01	paracetamol	3	g	O	
		3	g	P	
		3	g	R	

https://atcddd.fhi.no/atc_ddd_index/?code=N02BE01

Terminology Systems

Terminology Systems

Terminologies

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

LOINC

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

LOINC

q: paracetamol

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

LOINC

q: paracetamol *49* results

Terminology Systems

Terminologies

— vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

LOINC

q: paracetamol *49* results

4105-3	Acetaminophen [Mass] of Dose	Acetaminophen
60048-6	Acetaminophen [Mass/volume] in Blood	Acetaminophen
60049-4	Acetaminophen [Mass/volume] in Body fluid	Acetaminophen
35700-4	Acetaminophen [Mass/volume] in Gastric fluid	Acetaminophen
3298-7	Acetaminophen [Mass/volume] in Serum or Plasma	Acetaminophen
35595-8	Acetaminophen [Mass/volume] in Serum or Plasma by Screen method	Acetaminophen
13622-6	Acetaminophen [Mass/volume] in Urine	Acetaminophen
12442-0	Acetaminophen [Mass/volume] in Specimen	Acetaminophen
60050-2	Acetaminophen [Mass/mass] in Tissue	Acetaminophen
31089-6	Acetaminophen [Presence] in Bile fluid	Acetaminophen
88059-1	Acetaminophen [Presence] in Blood by Screen method	Acetaminophen
18265-9	Acetaminophen [Presence] in Body fluid	Acetaminophen
29389-4	Acetaminophen [Presence] in Gastric fluid	Acetaminophen

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

LOINC

q: paracetamol *49* results

4105-3	Acetaminophen [Mass] of Dose	Acetaminophen
60048-6	Acetaminophen [Mass/volume] in Blood	Acetaminophen
60049-4	Acetaminophen [Mass/volume] in Body fluid	Acetaminophen
35700-4	Acetaminophen [Mass/volume] in Gastric fluid	Acetaminophen
3298-7	Acetaminophen [Mass/volume] in Serum or Plasma	Acetaminophen
35595-8	Acetaminophen [Mass/volume] in Serum or Plasma by Screen method	Acetaminophen
13622-6	Acetaminophen [Mass/volume] in Urine	Acetaminophen
12442-0	Acetaminophen [Mass/volume] in Specimen	Acetaminophen
60050-2	Acetaminophen [Mass/mass] in Tissue	Acetaminophen
31089-6	Acetaminophen [Presence] in Bile fluid	Acetaminophen
88059-1	Acetaminophen [Presence] in Blood by Screen method	Acetaminophen
18265-9	Acetaminophen [Presence] in Body fluid	Acetaminophen
29389-4	Acetaminophen [Presence] in Gastric fluid	Acetaminophen

Terminology Systems

Terminologies

— vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

LOINC

q: paracetamol *49* results

Part Name	Acetaminophen
Part Display Name	Acetaminophen
Part Type	Component (Describes the core component or analyte measured)
Created On	2000-05-04
Construct for LOINC Short Name	APAP

Code: LP14712-1

4105-3	Acetaminophen [Mass] of Dose	Acetaminophen
60048-6	Acetaminophen [Mass/volume] in Blood	Acetaminophen
60049-4	Acetaminophen [Mass/volume] in Body fluid	Acetaminophen
35700-4	Acetaminophen [Mass/volume] in Gastric fluid	Acetaminophen
3298-7	Acetaminophen [Mass/volume] in Serum or Plasma	Acetaminophen
35595-8	Acetaminophen [Mass/volume] in Serum or Plasma by Screen method	Acetaminophen
13622-6	Acetaminophen [Mass/volume] in Urine	Acetaminophen
12442-0	Acetaminophen [Mass/volume] in Specimen	Acetaminophen
60050-2	Acetaminophen [Mass/mass] in Tissue	Acetaminophen
31089-6	Acetaminophen [Presence] in Bile fluid	Acetaminophen
88059-1	Acetaminophen [Presence] in Blood by Screen method	Acetaminophen
18265-9	Acetaminophen [Presence] in Body fluid	Acetaminophen
29389-4	Acetaminophen [Presence] in Gastric fluid	Acetaminophen

https://loinc.org/LP14712-1
https://loinc.org/search?t=1&l=en_US&q=paracetamol+

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

SNOMED-CT

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

SNOMED-CT

q: paracetamol

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

SNOMED-CT

q: paracetamol *518* results

Terminology Systems

Terminologies — vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

SNOMED-CT

q: paracetamol *518* results

Paracetamol (substance)
SCTID: 387517004

387517004 | Paracetamol (substance) |

en Paracetamol (substance)
en Acetaminophen
en Paracetamol

Terminology Systems

Terminologies

— vocabularies for specialised domains, concept-oriented, captures the relationship between concepts: **LOINC** (lab tests), **SNOMED-CT** (findings, disorders, procedures, etc)

SNOMED-CT

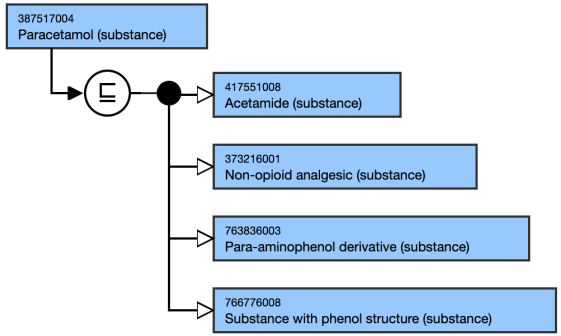
q: paracetamol *518* results

Paracetamol (substance)

SCTID: 387517004

387517004 | Paracetamol (substance) |

en Paracetamol (substance)
en Acetaminophen
en Paracetamol



https://snomedbrowser.org/?perspective=full&conceptId1=387517004&edition=MAIN/2026-08-01&release=&languages=en

Terminology Systems

Classifications

Terminology Systems

Classifications — arranges concepts into a set of categories according to explicit criteria: **ICD-10 / ICD-11** (diseases)

Terminology Systems

Classifications

— arranges concepts into a set of categories according to explicit criteria: **ICD-10 / ICD-11** (diseases)

- **ICD-10 Version:2019**
  - **I** Certain infectious and parasitic diseases
  - **II** Neoplasms
  - **III** Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
  - **IV** Endocrine, nutritional and metabolic diseases
  - **V** Mental and behavioural disorders
  - **VI** Diseases of the nervous system
  - **VII** Diseases of the eye and adnexa
  - **VIII** Diseases of the ear and mastoid process
  - **IX** Diseases of the circulatory system
  - **X** Diseases of the respiratory system
  - **XI** Diseases of the digestive system
  - **XII** Diseases of the skin and subcutaneous tissue
  - **XIII** Diseases of the musculoskeletal system and connective tissue
  - **XIV** Diseases of the genitourinary system
  - **XV** Pregnancy, childbirth and the puerperium

Terminology Systems

Classifications

— arranges concepts into a set of categories according to explicit criteria: **ICD-10 / ICD-11** (diseases)

- **ICD-10 Version:2019**
  - **I** Certain infectious and parasitic diseases
  - **II** Neoplasms
  - **III** Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
  - **IV** Endocrine, nutritional and metabolic diseases
  - **V** Mental and behavioural disorders
  - **VI** Diseases of the nervous system
  - **VII** Diseases of the eye and adnexa
  - **VIII** Diseases of the ear and mastoid process
  - **IX** Diseases of the circulatory system
  - **X** Diseases of the respiratory system
  - **XI** Diseases of the digestive system
  - **XII** Diseases of the skin and subcutaneous tissue
  - **XIII** Diseases of the musculoskeletal system and connective tissue
  - **XIV** Diseases of the genitourinary system
  - **XV** Pregnancy, childbirth and the puerperium

q: paracetamol

Terminology Systems

Classifications

— arranges concepts into a set of categories according to explicit criteria: **ICD-10 / ICD-11** (diseases)

- **ICD-10 Version:2019**
  - I Certain infectious and parasitic diseases
  - II Neoplasms
  - III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
  - IV Endocrine, nutritional and metabolic diseases
  - V Mental and behavioural disorders
  - VI Diseases of the nervous system
  - VII Diseases of the eye and adnexa
  - VIII Diseases of the ear and mastoid process
  - IX Diseases of the circulatory system
  - X Diseases of the respiratory system
  - XI Diseases of the digestive system
  - XII Diseases of the skin and subcutaneous tissue
  - XIII Diseases of the musculoskeletal system and connective tissue
  - XIV Diseases of the genitourinary system
  - XV Pregnancy, childbirth and the puerperium

q: paracetamol

paracetamol [Advanced Search] ICD-10 Versi

X40 Accidental poisoning by and exposure to nonopioid analgesics, antipyretics and antirheumatics
Paracetamol|Poisoning Accidental X40

Y10 Poisoning by and exposure to nonopioid analgesics, antipyretics and antirheumatics, undetermined intent
Paracetamol|Poisoning Undetermined intent Y10

X60 Intentional self-poisoning by and exposure to nonopioid analgesics, antipyretics and antirheumatics
Paracetamol|Poisoning Intentional self-harm X60

T39.1 Poisoning: 4-Aminophenol derivatives
Paracetamol|Poisoning Chapter 19 T39.1

Y45.5 4-Aminophenol derivatives
Paracetamol|Adverse effect in therapeutic use Y45.5

https://icd.who.int/browse10/2019/en#/T39.1

Terminology Systems

Terminology Systems

METATHESAURUS

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System)
integrates **195** **Terminology systems**

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates **195** **Terminology systems**

- links concepts through a shared **CUI** (Concept Unique Identifier)

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates **195** **Terminology systems**

- links concepts through a shared **CUI** (Concept Unique Identifier)
- the same concept in RxNorm, SNOMED, MeSH, etc. can resolve to one CUI

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates **195** **Terminology systems**

- links concepts through a shared **CUI** (Concept Unique Identifier)
- the same concept in RxNorm, SNOMED, MeSH, etc. can resolve to one CUI

- Thesauri, e.g., MeSH
- Statistical Classifications, e.g., ICD-9, ICD-10, ICPC
- Billing Codes, e.g., CPT, CPT Spanish version, HCPCS
- Clinical Coding Systems, e.g., SNOMED, Read
- Nursing Vocabularies, e.g., NIC, NOC, OMAHA
- Alternative/Complementary Medicine: ALTLINK
- Drug Sources: Multum, Micromedex, VANDF
- Drug Regulatory, e.g., MedDRA
- Lists of controlled terms, e.g., COSTAR, HL7 Arden Syntax

Categories of source vocabulary integrated into the UMLS

Dogac, Asuman. "Interoperability in eHealth systems." *Proceedings of the VLDB Endowment* 5.12 (2012): 2026-2027.

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates

q: paracetamol

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates

q: paracetamol

acetaminophen

UMLS CUI: C0000970

Semantic Types: Organic Chemical · Pharmacologic Substance

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates

q: paracetamol

acetaminophen

UMLS CUI: C0000970

Semantic Types: Organic Chemical · Pharmacologic Substance

Terminology Systems

METATHESAURUS

UMLS (Unified Medical Language System) integrates

q: paracetamol

acetaminophen

UMLS CUI: C0000970

Semantic Types: Organic Chemical · Pharmacologic Substance

Name	AUI	Vocabulary	Term Type	Code
acetaminophen	A31525798	MTH	PN	NOCODE
acetaminophen	A31703571	RXNORM	IN	161
Acetaminophen	A0018250	MSH	MH	D000082
Acetamidophenol	A26662484	MSH	ET	D000082
Acetominophen	A26654540	MSH	ET	D000082
Hydroxyacetanilide	A26638476	MSH	ET	D000082
p-Acetamidophenol	A26608671	MSH	ET	D000082
p-Hydroxyacetanilide	A26600498	MSH	ET	D000082
Paracetamol	A26627559	MSH	ET	D000082
N-Acetyl-p-aminophenol	A26622126	MSH	ET	D000082
N-(4-Hydroxyphenyl)acetanilide	A26597837	MSH	ET	D000082
APAP	A26654541	MSH	ET	D000082
Acetaminophen-containing product	A30195893	SNOMEDCT_US	PT	90332006
Acetaminophen	A2878027	SNOMEDCT_US	PT	387517004
Paracetamol (substance)	A28464918	SNOMEDCT_US	FN	387517004
Product containing paracetamol (medicinal product)	A28435625	SNOMEDCT_US	FN	90332006
Paracetamol	A28410045	SNOMEDCT_US	SY	387517004
Paracetamol-containing product	A30195894	SNOMEDCT_US	SY	90332006

https://uts.nlm.nih.gov/uts/umls/concept/C0000970

Data Harmonisation Challenges

Data Harmonisation Challenges

Data integration

Data Harmonisation Challenges

Data integration

Data harmonisation

Data Harmonisation Challenges

Data integration — providing a unified, queryable view over sources.

Data harmonisation

Data Harmonisation Challenges

Data integration — providing a unified, queryable view over sources.

Data harmonisation — mapping local content into a common representation, reconciling representations so values mean the same thing and can be pooled.

Data Harmonisation Challenges

Data integration — providing a unified, queryable view over sources.

Data harmonisation — mapping local content into a common representation, reconciling representations so values mean the same thing and can be pooled.

THE CHALLENGE

Data Harmonisation Challenges

Data integration — providing a unified, queryable view over sources.

Data harmonisation — mapping local content into a common representation, reconciling representations so values mean the same thing and can be pooled.

THE CHALLENGE

Concept normalisation — mapping a local code/string to a concept in a reference vocabulary (e.g., SNOMED, LOINC)

Data Harmonisation Challenges

Data integration — providing a unified, queryable view over sources.

Data harmonisation — mapping local content into a common representation, reconciling representations so values mean the same thing and can be pooled.

THE CHALLENGE

Concept normalisation — mapping a local code/string to a concept in a reference vocabulary (e.g., SNOMED, LOINC)

ALSO KNOWN AS …

Data Harmonisation Challenges

Not one step, but **a combination of**:

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

Cell value to KG

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction

Cell value to KG

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction

Cell value to KG

- The entity is already “extracted”

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction
- Entity Linking / (Named)-Entity Disambiguation

Cell value to KG

- The entity is already “extracted”

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction
- Entity Linking / (Named)-Entity Disambiguation

Cell value to KG

- The entity is already “extracted”
- Semantic Table Interpretation / Cell-entity Annotation, SemTab Challenge

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction
- Entity Linking / (Named)-Entity Disambiguation

Cell value to KG

- The entity is already “extracted”
- Semantic Table Interpretation / Cell-entity Annotation, SemTab Challenge

Limaye et al. 2010 (graphical model based), Ritze et al. 2014 (T2K Match), Efthymiou et al. 2017 (Matching Web Tables), Deng et al. 2022 (Turl), Ngyuen et al. 2021 (Mtab), Dong et al. 2023 (BLINKout), Shen et al. 2015 “Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions”, Sevgili et al. 2022 “Neural Entity Linking: A Survey of Models Based on Deep Learning”

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction
- Entity Linking / (Named)-Entity Disambiguation

Cell value to KG

- The entity is already “extracted”
- Semantic Table Interpretation / Cell-entity Annotation, SemTab Challenge

YES, BUT …

Limaye et al. 2010 (graphical model based), Ritze et al. 2014 (T2K Match), Efthymiou et al. 2017 (Matching Web Tables), Deng et al. 2022 (Turl), Ngyuen et al. 2021 (Mtab), Dong et al. 2023 (BLINKout), Shen et al. 2015 “Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions”, Sevgili et al. 2022 “Neural Entity Linking: A Survey of Models Based on Deep Learning”

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction
- Entity Linking / (Named)-Entity Disambiguation

Cell value to KG

- The entity is already "extracted"
- Semantic Table Interpretation / Cell-entity Annotation, SemTab Challenge

YES, BUT …

1. Characterise ambiguity: know when *not* to map (e.g., *cold*), KG are not complete

Limaye et al. 2010 (graphical model based), Ritze et al. 2014 (T2K Match), Efthymiou et al. 2017 (Matching Web Tables), Deng et al. 2022 (Turl), Ngyuen et al. 2021 (Mtab), Dong et al. 2023 (BLINKout), Shen et al. 2015 "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions", Sevgili et al. 2022 "Neural Entity Linking: A Survey of Models Based on Deep Learning"

Data Harmonisation Challenges

Not one step, but **a combination of**:

Text mention to KG

- First step: entity extraction
- Entity Linking / (Named)-Entity Disambiguation

Cell value to KG

- The entity is already "extracted"
- Semantic Table Interpretation / Cell-entity Annotation, SemTab Challenge

YES, BUT …

- Characterise ambiguity: know when *not* to map (e.g., *cold*), KG are not complete
- Domain specific representations (e.g., abbreviations)

Limaye et al. 2010 (graphical model based), Ritze et al. 2014 (T2K Match), Efthymiou et al. 2017 (Matching Web Tables), Deng et al. 2022 (Turl), Ngyuen et al. 2021 (Mtab), Dong et al. 2023 (BLINKout), Shen et al. 2015 "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions", Sevgili et al. 2022 "Neural Entity Linking: A Survey of Models Based on Deep Learning"

Ontology Alignment Challenges

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Terminology systems describe the world differently, so alignment is lossy or ambiguous

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Terminology systems describe the world differently, so alignment is lossy or ambiguous

THE SOLUTION

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Terminology systems describe the world differently, so alignment is lossy or ambiguous

THE SOLUTION

Ontology mapping and repair tools: **LogMap**, **AgreementMakerLight**, **BERTMap**

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Terminology systems describe the world differently, so alignment is lossy or ambiguous

THE SOLUTION

Ontology mapping and repair tools: **LogMap**, **AgreementMakerLight**, **BERTMap**

Jiménez-Ruiz & Cuenca Grau. 2011 (LogMap), Faria et al. 2013 (AgreementMakerLight), He at al. 2022 (BERTMap), Thieblin et al. 2020 “Survey on complex ontology matching”

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Terminology systems describe the world differently, so alignment is lossy or ambiguous

THE SOLUTION

Ontology mapping and repair tools: **LogMap**, **AgreementMakerLight**, **BERTMap**

YES, BUT …

Jiménez-Ruiz & Cuenca Grau. 2011 (LogMap), Faria et al. 2013 (AgreementMakerLight), He at al. 2022 (BERTMap), Thieblin et al. 2020 “Survey on complex ontology matching”

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Terminology systems describe the world differently, so alignment is lossy or ambiguous

THE SOLUTION

Ontology mapping and repair tools: **LogMap**, **AgreementMakerLight**, **BERTMap**

YES, BUT …

1. Foundational mismatch: linking classifications (ICD-10) with vocabularies (SNOMED) requires a shared model of meaning (e.g., BioTop, BFO2), but no one formalism suffices

Jiménez-Ruiz & Cuenca Grau. 2011 (LogMap), Faria et al. 2013 (AgreementMakerLight), He at al. 2022 (BERTMap), Thieblin et al. 2020 “Survey on complex ontology matching”

Ontology Alignment Challenges

So far we assumed a *single* target — map to SNOMED, ICD, or LOINC. But what about reconciling ontologies directly, e.g., SNOMED to ICD?

THE CHALLENGE

Terminology systems describe the world differently, so alignment is lossy or ambiguous

THE SOLUTION

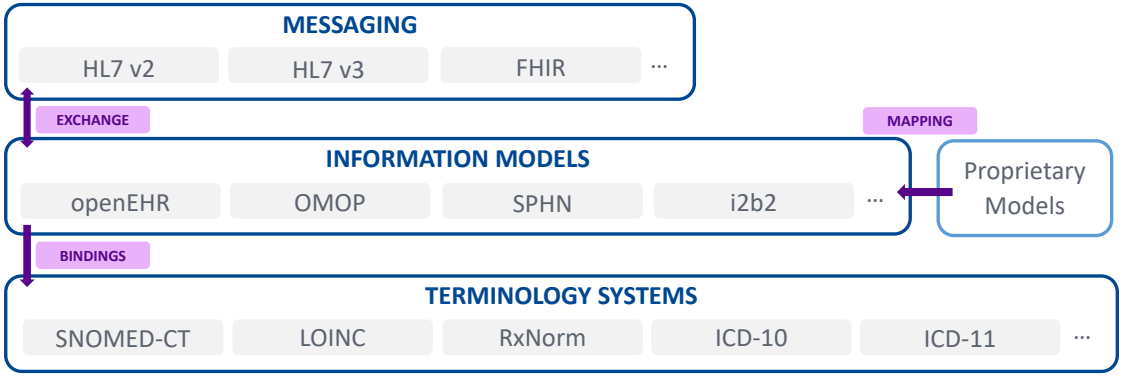
Ontology mapping and repair tools: **LogMap**, **AgreementMakerLight**, **BERTMap**

YES, BUT …

- Foundational mismatch: linking classifications (ICD-10) with vocabularies (SNOMED) requires a shared model of meaning (e.g., BioTop, BFO2), but no one formalism suffices
- Complex correspondences: 1-n, n-m, and subsumption relationships

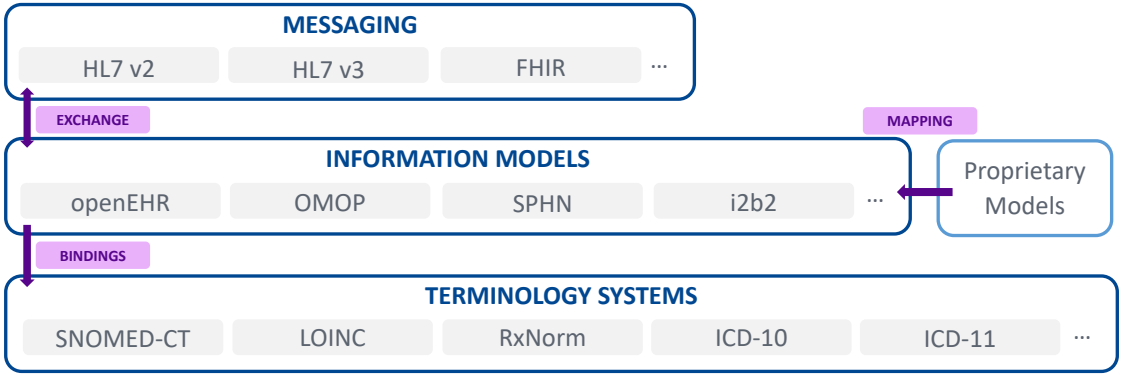
Jiménez-Ruiz & Cuenca Grau. 2011 (LogMap), Faria et al. 2013 (AgreementMakerLight), He at al. 2022 (BERTMap), Thieblin et al. 2020 “Survey on complex ontology matching”

Constructing a 2nd Gen Pipeline



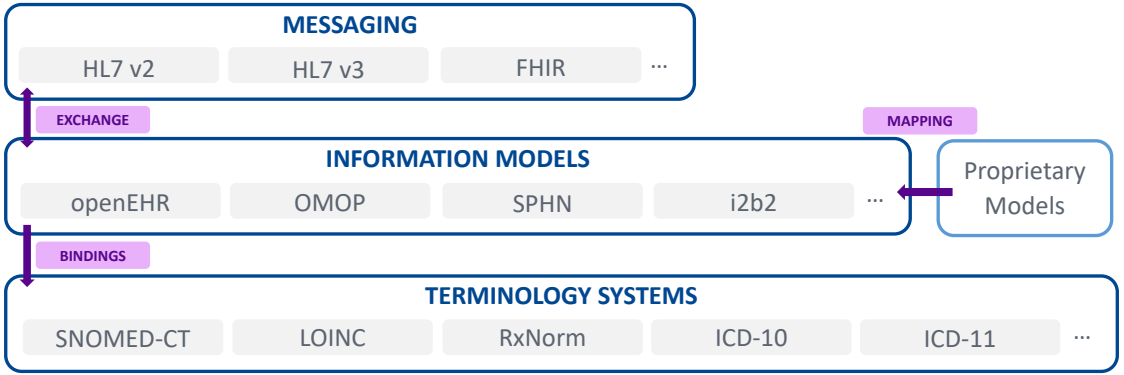
Constructing a 2nd Gen Pipeline

Governing bodies decide the “interoperability stack”



Constructing a 2nd Gen Pipeline

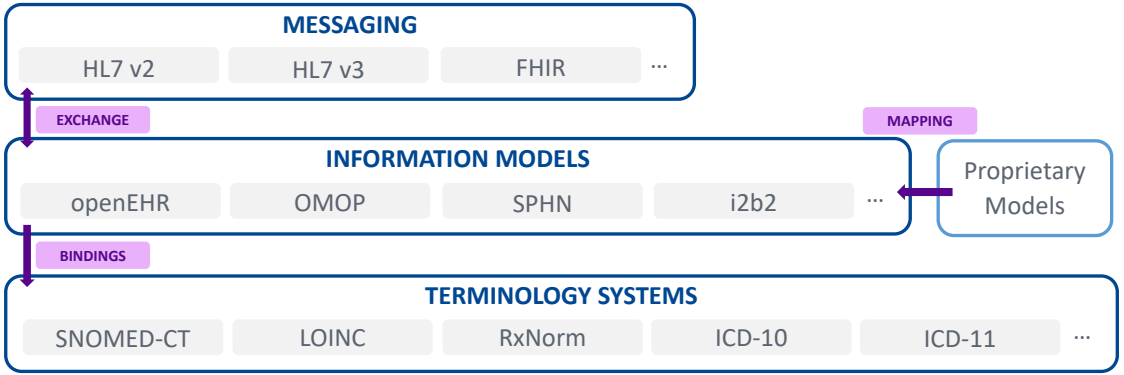
Governing bodies decide the “interoperability stack”



EXAMPLES

Constructing a 2nd Gen Pipeline

Governing bodies decide the “interoperability stack”



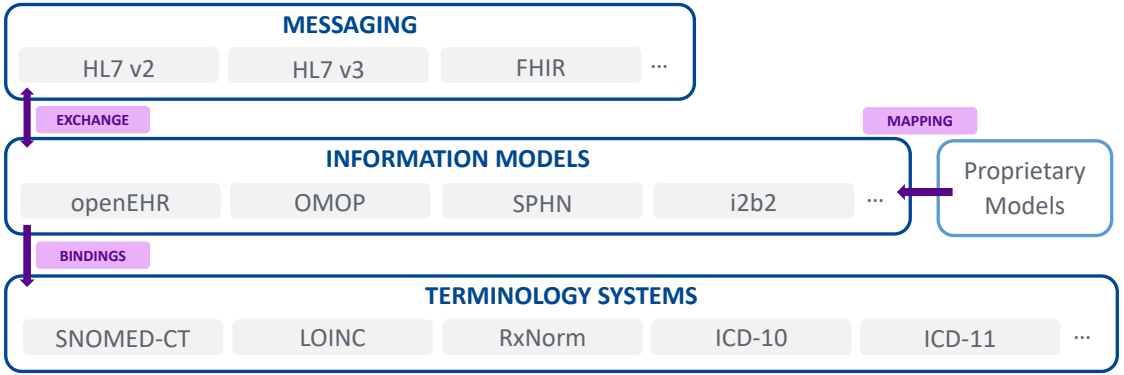
EXAMPLES

EU

HL7 FHIR + IHE profiles + SNOMED CT, LOINC

Constructing a 2nd Gen Pipeline

Governing bodies decide the “interoperability stack”



EXAMPLES

EU

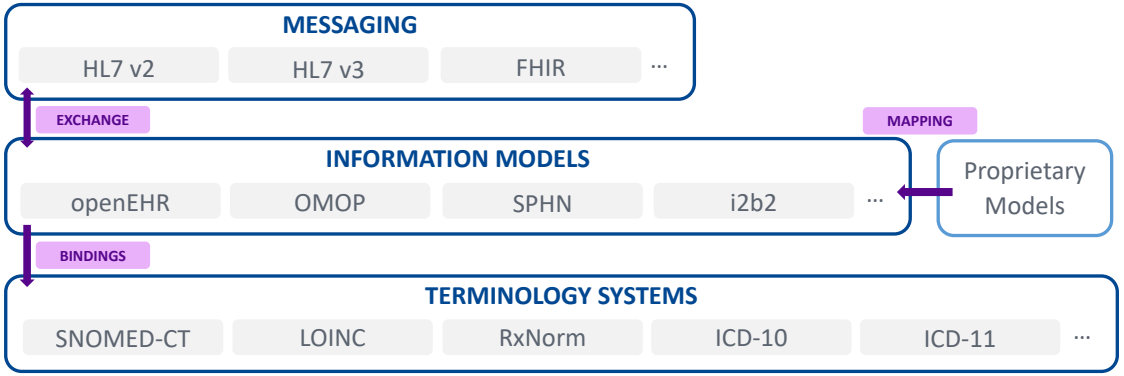
HL7 FHIR + IHE profiles + SNOMED CT, LOINC

USA

HL7 FHIR + CDA + SNOMED CT, RxNORM, LOINC, ICD-10, CVX

Constructing a 2nd Gen Pipeline

Governing bodies decide the “interoperability stack”



PROBLEM

EXAMPLES

EU

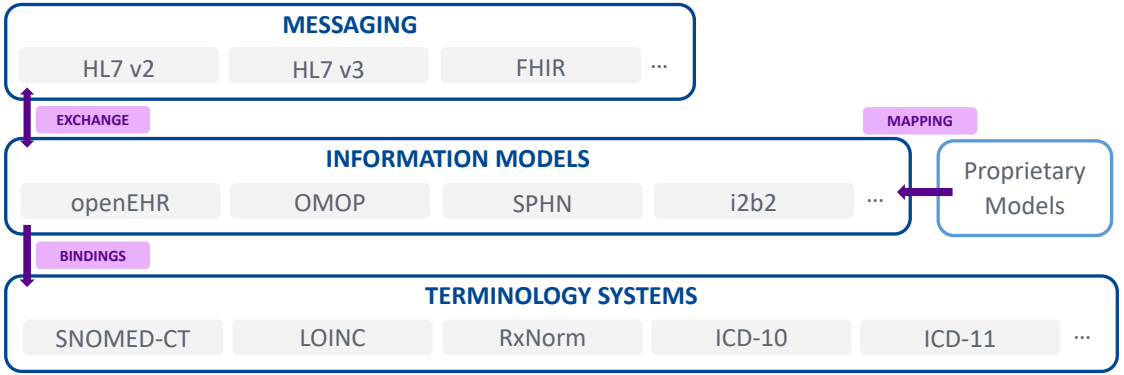
HL7 FHIR + IHE profiles + SNOMED CT, LOINC

USA

HL7 FHIR + CDA + SNOMED CT, RxNORM, LOINC, ICD-10, CVX

Constructing a 2nd Gen Pipeline

Governing bodies decide the “interoperability stack”



EXAMPLES

EU

HL7 FHIR + IHE profiles + SNOMED CT, LOINC

USA

HL7 FHIR + CDA + SNOMED CT, RxNORM, LOINC, ICD-10, CVX

PROBLEM

- Many semantic traps: losing the meaning from mapping to mapping

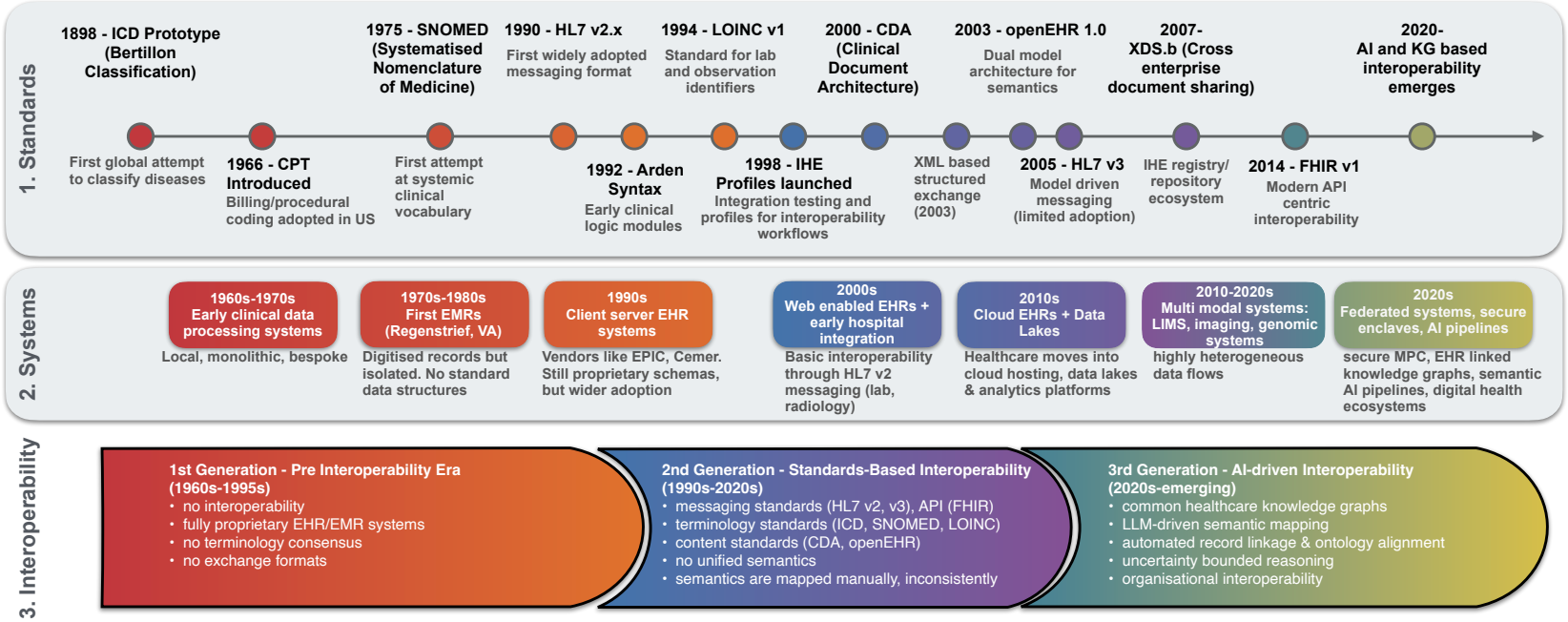
PART 03

Third Generation

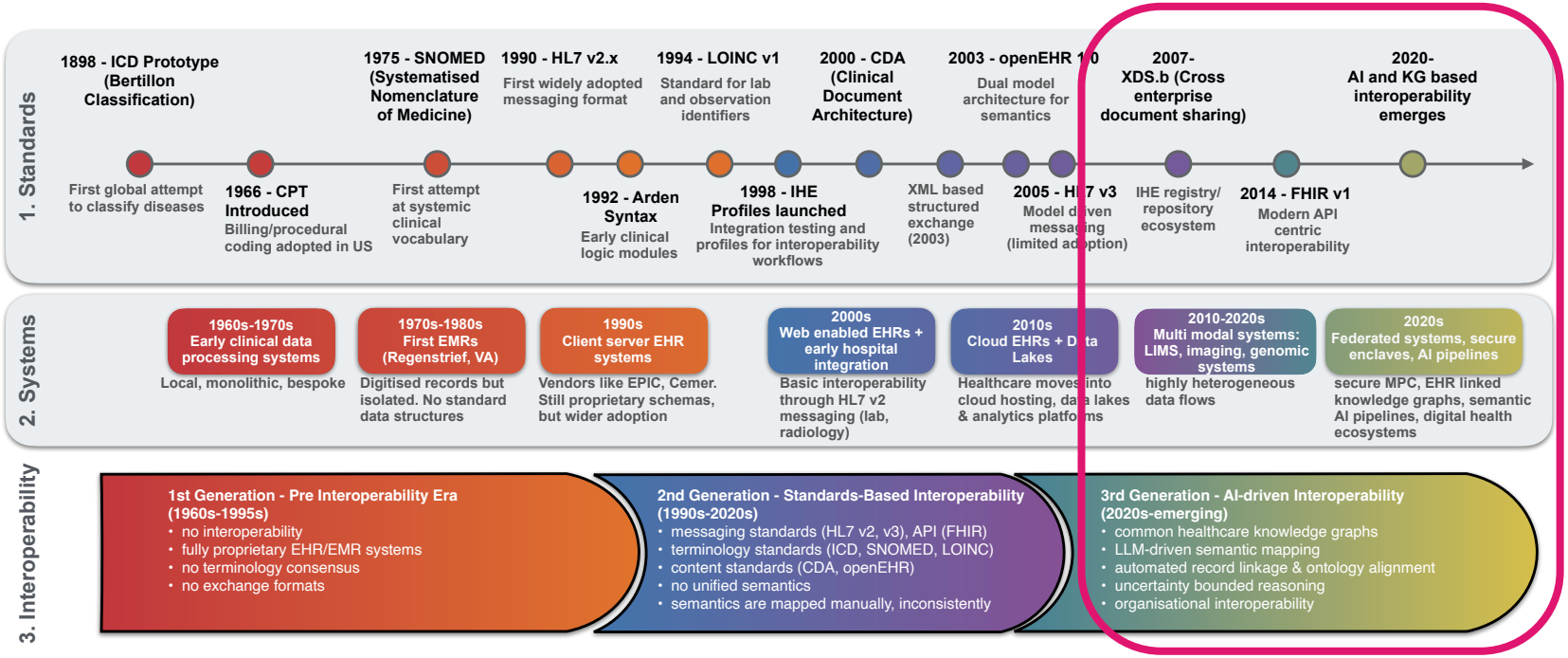
Katja Hose, Paris Carbone, Christos Koutras

KTH / NYU / TU Wien

History


1. Standards
1898 - ICD Prototype (Bertillon Classification)
First global attempt to classify diseases
1966 - CPT Introduced
Billing/procedural coding adopted in US
1975 - SNOMED (Systematised Nomenclature of Medicine)
First attempt at systemic clinical vocabulary
1990 - HL7 v2.x
First widely adopted messaging format
1992 - Arden Syntax
Early clinical logic modules
1994 - LOINC v1
Standard for lab and observation identifiers
1998 - IHE Profiles launched
Integration testing and profiles for interoperability workflows
2000 - CDA (Clinical Document Architecture)
XML based structured exchange (2003)
2003 - openEHR 1.0
Dual model architecture for semantics
2005 - HL7 v3
Model driven messaging (limited adoption)
2007- XDS.b (Cross enterprise document sharing)
IHE registry/ repository ecosystem
2014 - FHIR v1
Modern API centric interoperability
2020- AI and KG based interoperability emerges
2. Systems
1960s-1970s Early clinical data processing systems
Local, monolithic, bespoke
1970s-1980s First EMRs (Regenstrief, VA)
Digitised records but isolated. No standard data structures
1990s Client server EHR systems
Vendors like EPIC, Cemer. Still proprietary schemas, but wider adoption
2000s Web enabled EHRs + early hospital integration
Basic interoperability through HL7 v2 messaging (lab, radiology)
2010s Cloud EHRs + Data Lakes
Healthcare moves into cloud hosting, data lakes & analytics platforms
2010-2020s Multi modal systems: LIMS, imaging, genomic systems
highly heterogeneous data flows
2020s Federated systems, secure enclaves, AI pipelines
secure MPC, EHR linked knowledge graphs, semantic AI pipelines, digital health ecosystems
3. Interoperability
1st Generation - Pre Interoperability Era (1960s-1995s)
• no interoperability
• fully proprietary EHR/EMR systems
• no terminology consensus
• no exchange formats
2nd Generation - Standards-Based Interoperability (1990s-2020s)
• messaging standards (HL7 v2, v3), API (FHIR)
• terminology standards (ICD, SNOMED, LOINC)
• content standards (CDA, openEHR)
• no unified semantics
• semantics are mapped manually, inconsistently
3rd Generation - AI-driven Interoperability (2020s-emerging)
• common healthcare knowledge graphs
• LLM-driven semantic mapping
• automated record linkage & ontology alignment
• uncertainty bounded reasoning
• organisational interoperability

History


1. Standards
1898 - ICD Prototype (Bertillon Classification)
First global attempt to classify diseases
1966 - CPT Introduced
Billing/procedural coding adopted in US
1975 - SNOMED (Systematised Nomenclature of Medicine)
First attempt at systemic clinical vocabulary
1990 - HL7 v2.x
First widely adopted messaging format
1992 - Arden Syntax
Early clinical logic modules
1994 - LOINC v1
Standard for lab and observation identifiers
1998 - IHE Profiles launched
Integration testing and profiles for interoperability workflows
2000 - CDA (Clinical Document Architecture)
XML based structured exchange (2003)
2003 - openEHR 1.0
Dual model architecture for semantics
2005 - HL7 v3
Model driven messaging (limited adoption)
2007- XDS.b (Cross enterprise document sharing)
IHE registry/ repository ecosystem
2014 - FHIR v1
Modern API centric interoperability
2020- AI and KG based interoperability emerges
2. Systems
1960s-1970s Early clinical data processing systems
Local, monolithic, bespoke
1970s-1980s First EMRs (Regenstrief, VA)
Digitised records but isolated. No standard data structures
1990s Client server EHR systems
Vendors like EPIC, Cemer. Still proprietary schemas, but wider adoption
2000s Web enabled EHRs + early hospital integration
Basic interoperability through HL7 v2 messaging (lab, radiology)
2010s Cloud EHRs + Data Lakes
Healthcare moves into cloud hosting, data lakes & analytics platforms
2010-2020s Multi modal systems: LIMS, imaging, genomic systems
highly heterogeneous data flows
2020s Federated systems, secure enclaves, AI pipelines
secure MPC, EHR linked knowledge graphs, semantic AI pipelines, digital health ecosystems
3. Interoperability
1st Generation - Pre Interoperability Era (1960s-1995s)
• no interoperability
• fully proprietary EHR/EMR systems
• no terminology consensus
• no exchange formats
2nd Generation - Standards-Based Interoperability (1990s-2020s)
• messaging standards (HL7 v2, v3), API (FHIR)
• terminology standards (ICD, SNOMED, LOINC)
• content standards (CDA, openEHR)
• no unified semantics
• semantics are mapped manually, inconsistently
3rd Generation - AI-driven Interoperability (2020s-emerging)
• common healthcare knowledge graphs
• LLM-driven semantic mapping
• automated record linkage & ontology alignment
• uncertainty bounded reasoning
• organisational interoperability

Integrating Health Data using Graphs

Example Scenario

- Two hospitals contribute retrospective cohorts of patients
    - No patient appears in both, cannot be linked directly
- Heterogeneous representation of data at the hospitals (column names, codes, details , languages, etc.)
- Partial overlap of types of information between hospitals

Different Levels of Detail

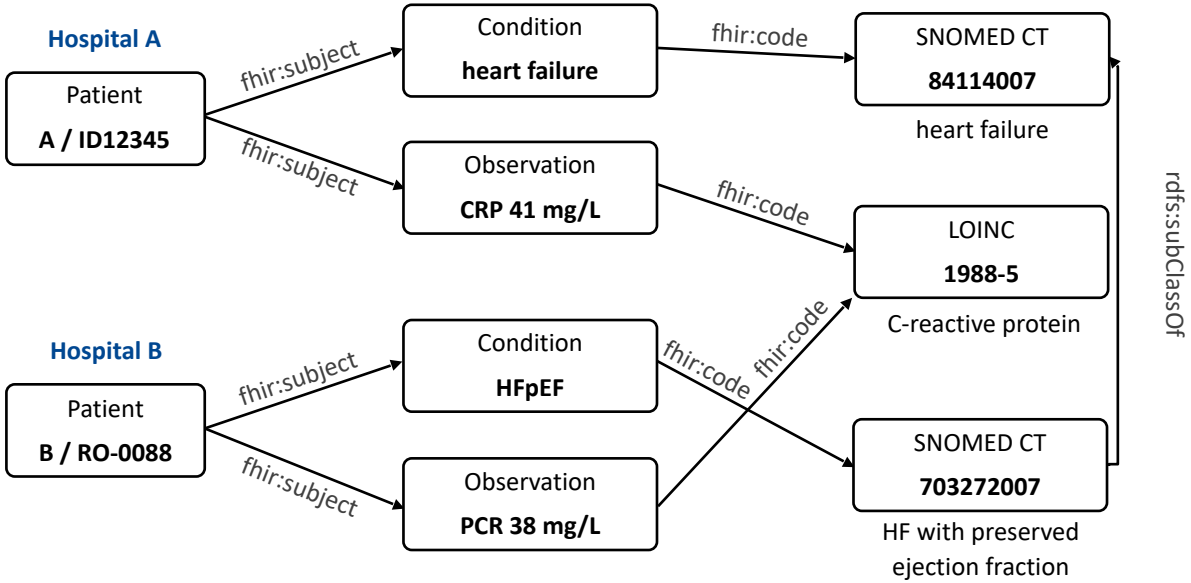
Example

	hospital A	hospital B
heart failure	prev_heart_failure	HFpEF / HFmrEF / HFrEF
antithrombotics	prev_antiplatelet, prev_anticoagulant	ASA / P2Y12 / AVK / DOAC / ...
C-reactive protein	crp	PCR

The two hospitals describe the same clinical concepts at different levels of detail

- One boolean (prev_heart_failure) vs. three
- Two high-level categories vs. sixteen named drug/drug-level classes, e.g., ASA – Aspirin
- Two different abbreviations denoting the same thing
- …

Integrating Heterogeneous Patients using Graphs



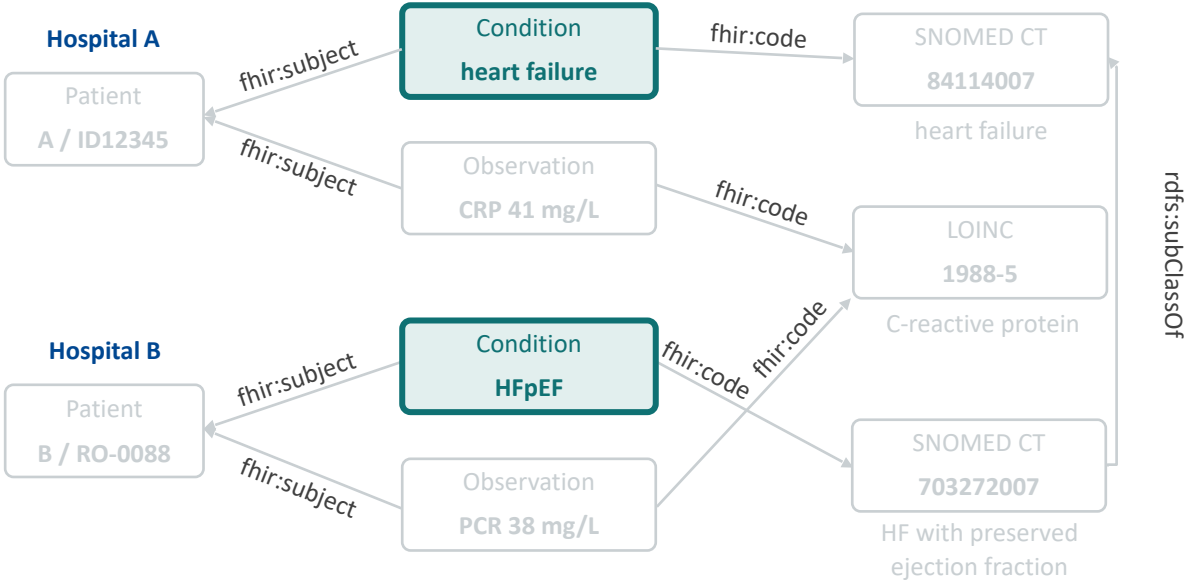
- No need to design a joint schema or harmonize data upfront
- Patient cohorts remain disjoint but become linked via concepts
- Examples
  - "crp" and "PCR" resolve to the same LOINC code
  - SNOMED links two heart-failure codes via a "subclassOf" edge

Queries for Retrieving Integrated Data

All patients with heart failure from all hospitals

```
SELECT ?patient WHERE {
  ?cond rdf:type fhir:Condition . # find all conditions
  ?cond fhir:subject ?patient .   # find the patients with the identified conditions
  ?cond fhir:code ?code .         # find the code of the identified conditions
  ?code rdfs:subClassOf* <http://snomed.info/id/84114007> .     # restrict to codes that
                                                          # heart failure, incl. any subtype
                                                          # of it
}
```

Match the Condition Pattern

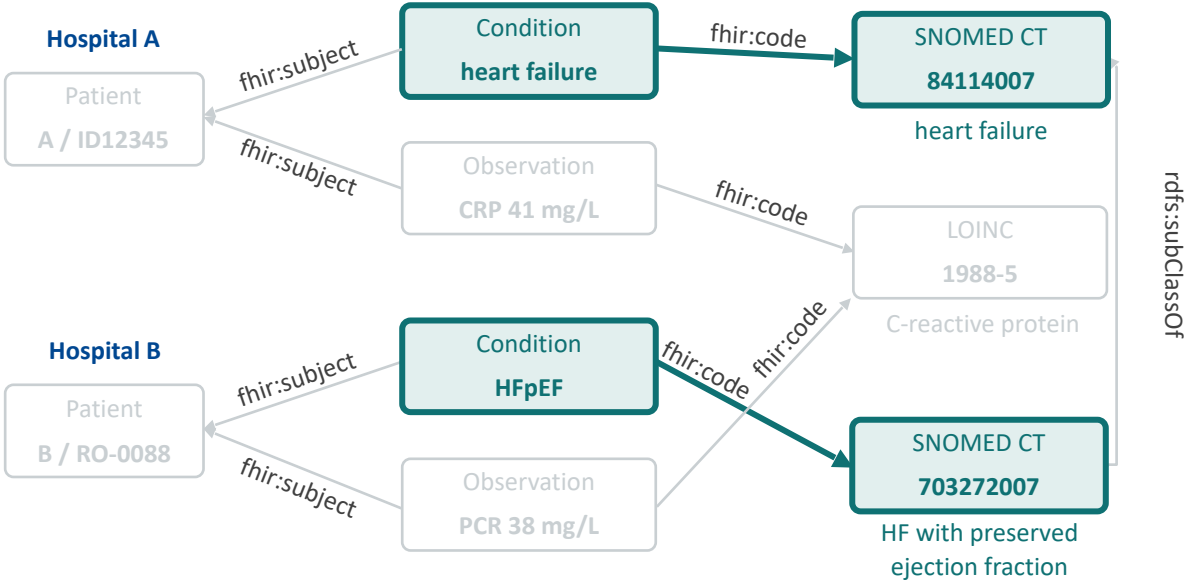


Step 1 of 5

```
?cond rdf:type fhir:Condition
```

- Find all nodes of type Condition in our graph

Retrieve the Codes for the Conditions

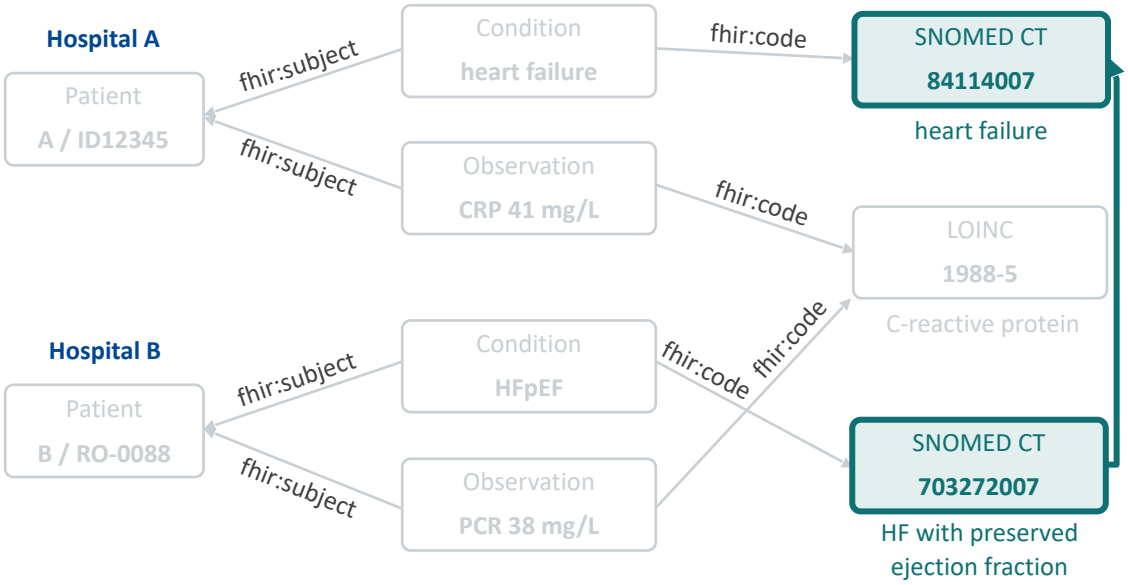


Step 2 of 5

```
?cond fhir:code ?code
```

- Each condition carries a code
- Hospitals use different codes/ concepts for heart failures: 84114007 and 703272007

Traversing the Hierarchy

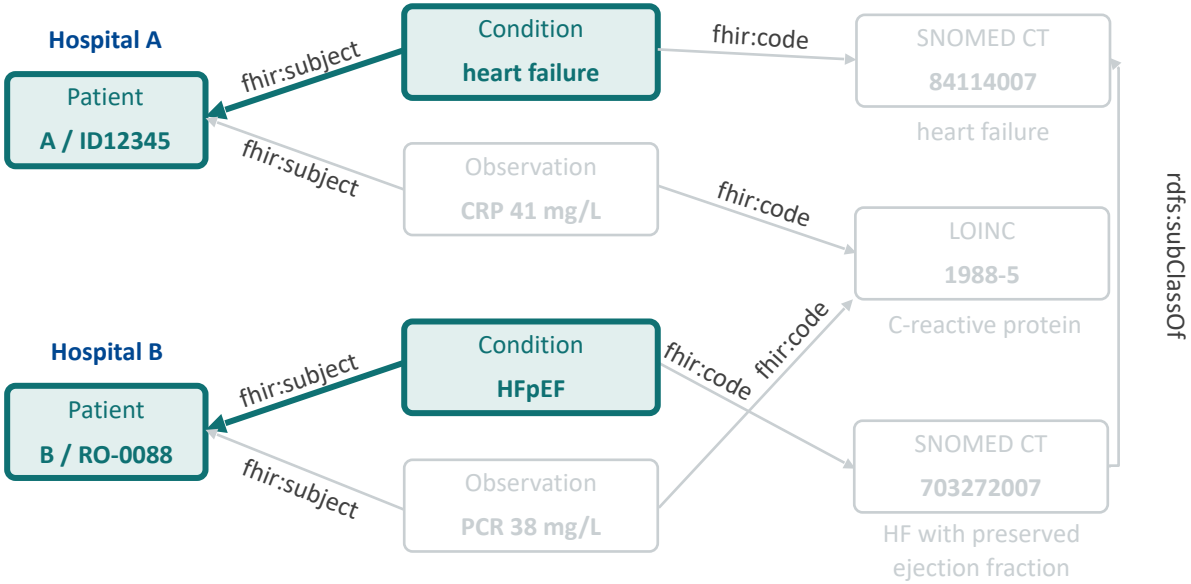


Step 3 of 5

```
?code rdfs:subClassOf*
    sct:84114007
```

- * matches zero hops for 84114007 itself, and one hop for 703272007
- This connection comes from SNOMED directly
- Both matched concepts are of type heart failure

Retrieving Patients with those Conditions

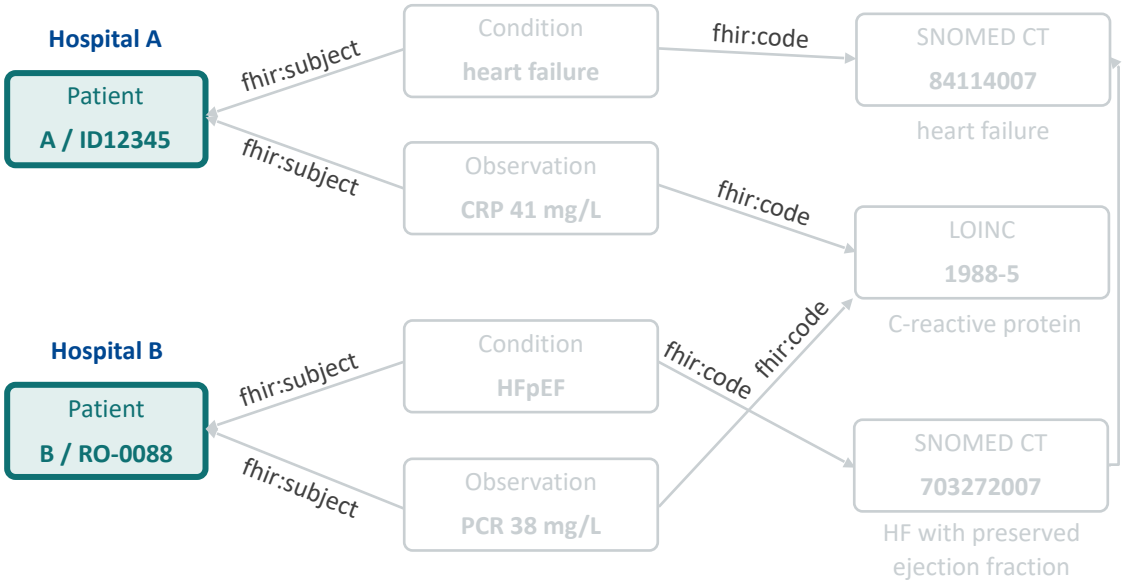


Step 4 of 5

```
?cond fhir:subject ?patient
```

- We have two patients linked to these conditions

The Result



Step 5 of 5

```
SELECT ?patient
```

- Output the nodes of the patients suffering from heart conditions
- Adding more patients only requires connecting them to standardized conditions

Actual Query is Slightly More Complicated

FHIR is more verbose and complex, so the actual FHIR-based query looks like this

```
SELECT ?patient WHERE {
  ?cond rdf:type fhir:Condition . # find all conditions
  ?cond fhir:subject/fhir:link ?patient . # the patient URI (ID) is nested inside another
  ?cond fhir:code/fhir:coding/fhir:code/fhir:v ?code . # the code is nested even deeper
  BIND( IRI(CONCAT("http://snomed.info/id/", ?code)) AS ?concept ) # the code is an
                                                          # integer or alphanumeric value,
                                                          # so we use BIND to transform it
                                                          # into a URI (URL)
  ?concept rdfs:subClassOf* <http://snomed.info/id/84114007> . # restrict to codes that re
                                                          # heart failure, incl. any subtype
                                                          # of it}
```

Graph-Based Integration

- Extensible to further hospitals with different representations
- External standards and resources for linking related concepts (LOINC, SNOMED, etc.)
- Ontologies can be integrated flexibly
- Flexible with respect to updates
- …

Graph-Based Integration

- Extensible to further hospitals with different representations
- External standards and resources for linking related concepts (LOINC, SNOMED, etc.)
- Ontologies can be integrated flexibly
- Flexible with respect to updates
- …

Challenges

- Mapping local codes to external standards
  Example: "crp" and "pcr" to "1988-5“ (LOINC)
- More computational load when retrieving the data
- …

Integration using LLMs: A Case Study

Denmark has a centralized national hospital register since the 70s (The Danish National Patient Register)

Evolving registries (three major revisions so far)

CPR
Det Centrale Personregister

```
t_person
t_person_status
t_adresse_udland
+2 history
```

LPR1 / LPR2
Landspatientregisteret, 1977–

```
t_adm
t_diag
t_sksube
+3 more
```

LPR2 psyk.
psychiatric extension, 1995–

```
t_psyk_adm
t_psyk_diag
t_psyk_sksopr
+4 more
```

LPR3
rebuilt 2019, SOR-based

```
kontakter
diagnoser
procedurer_kirurgi
+2 more
```

DAR
Dødsårsagsregisteret

```
t_dodsaarsag_1
t_dodsaarsag_2
```

LSR
Lægemiddelstatistikregisteret

```
epikur
laegemiddel
+1 more
```

SMR
Sygehusmedicinregisteret, 2018–

```
indberetningmedpris
```

LAB
Laboratoriedatabasen

```
lab_dm_forsker
lab_dm_labidcodes
+1 more
```

SOR
Sundhedsvæsenets Organisationsregister

```
organisationer
SOREntity
shakcomplete
```

Example:
SKS KTPJ05E
"Nålebiopsi af ingvinal lymfeknude" · 2,860 rows
→ **SNOMED 4120431**

Registries CPR Central Personal Registry · LPR National Patient Register · DAR Cause of Death Register · LSR Medicines Statistics Register · SMR Hospital Medicine Register · LAB Laboratory Database · SOR Healthcare Organization Register

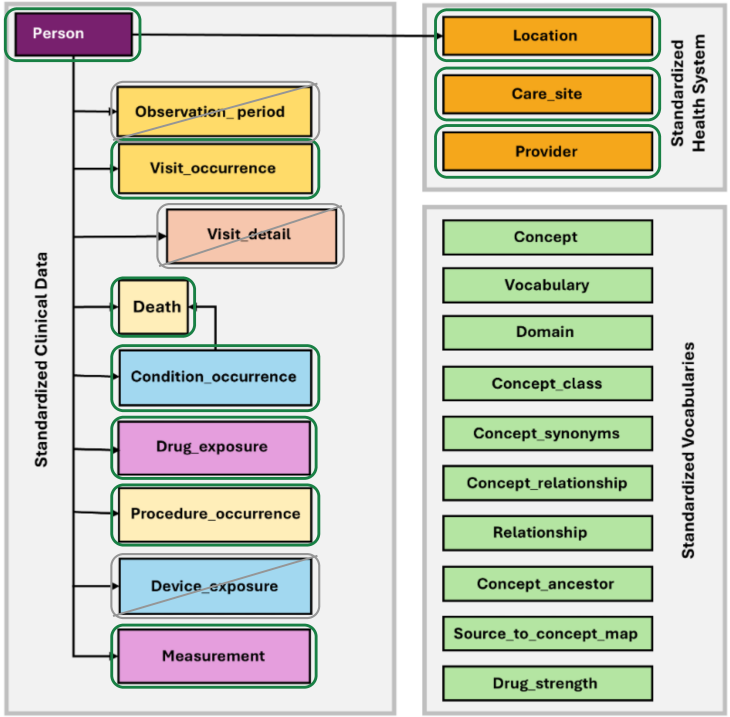
Code systems ICD International Classification of Diseases (-DK = Danish edition) · SKS Danish Healthcare Classification System · NPU Nomenclature for Properties and Units · ATC Anatomical Therapeutic Chemical · VNR Varenummer (product number) · SHAK Hospital Department Classification

Joint Analysis Requires Integration

Mapping to OMOP

OMOP table	From registries	Records
person	CPR	4.89 M
death	DAR + CPR	1.81 M
visit_occurrence	t_adm + kontakter	108.8 M
condition_occurrence	t_diag	251.5 M
procedure_occurrence	SKS + LPR3	34.6 M
drug_exposure	epikur + SMR	1,095 M
measurement	lab_dm_forsker	1,672 M
care_site + provider	SOR, SHAK, LAB	155 k
location	geo files	1 k

3.8 billion source records, 11 national registries, 9.1 million patients



Four Ways to Map Codes and Concepts

Which method is applicable depends on the concrete issue

Crosswalk

use when

the mapping already exists

ICD-10-DK, ATC, VNR

130,167 rows loaded as a dictionary

what we achieved

89.3% of 312.7 M condition rows
121,834 ICD-10-DK codes
99.93% of 1.34 B drug rows

Mapping Tool (Usagi)

use when

English terms, ranked suggestions

SKS, NPU

1,242 procedure mappings after expert review

what we achieved

93.35% of procedures
32.3 M of 34.6 M rows
99.24% of 1.59 B lab rows

Keyword rules

use when

naming is consistent

care sites: “apotek”, “tandpleje”

37,028 of 150,191 sites (24.7%)

what we achieved

37,028 of 150,191 sites
4,191 SOR codes inspected
113,163 left over

LLM

use when

Danish free text, no pattern, too big to map by hand

the remaining care sites

113,163 sites, $65.82

what we achieved

113,163 sites classified
77% accuracy, $65.82 total

ATC — Anatomical Therapeutic Chemical Classification Code
VNR — Varenummer — product/article number

SKS — Sundhedsvæsenets Klassifikationssystem — Clinical and Administrative Information
NPU — Nomenclature for Properties and Units
SOR — Sundhedsvæsenets Organisationsregister — Danish Healthcare Organization Register

Four Ways to Map Codes and Concepts

Which method is applicable depends on the concrete issue

Crosswalk

use when

the mapping already exists

ICD-10-DK, ATC, VNR

130,167 rows loaded as a dictionary

what we achieved

89.3% of 312.7 M condition rows
121,834 ICD-10-DK codes
99.93% of 1.34 B drug rows

Mapping Tool (Usagi)

use when

English terms, ranked suggestions

SKS, NPU

1,242 procedure mappings after expert review

what we achieved

93.35% of procedures
32.3 M of 34.6 M rows
99.24% of 1.59 B lab rows

Keyword rules

use when

naming is consistent

care sites: “apotek”, “tandpleje”

37,028 of 150,191 sites (24.7%)

what we achieved

37,028 of 150,191 sites
4,191 SOR codes inspected
113,163 left over

LLM

use when

Danish free text, no pattern, too big to map by hand

the remaining care sites

113,163 sites, $65.82

what we achieved

113,163 sites classified
77% accuracy, $65.82 total

ATC — Anatomical Therapeutic Chemical Classification Code
VNR — Varenummer — product/article number

SKS — Sundhedsvæsenets Klassifikationssystem — Clinical and Administrative Information
NPU — Nomenclature for Properties and Units
SOR — Sundhedsvæsenets Organisationsregister — Danish Healthcare Organization Register

Example Classification using LLMs

150,191
SOR organizations from the register

→

37,028
keyword rules and human review first

→

113,163
left for the LLM

Example prompt

```
You are a medical administrator. Classify each medical organization into the most appropriate standard category,
selecting the most specific type when more than one may apply.

Example Mappings
Nordhavn Apotek, filial af Farum Apotek  →  Pharmacy
OUH Neurokirurgisk Afsnit U  →  Inpatient Hospital
L. Tronegård & M. S. Schmidt  →  Ambulatory Primary Care Clinic
Hudlægeklinikken Frederikshavn  →  Ambulatory Medical Specialty Clinic
Psykologhuset Svenstrup  →  Ambulatory Adult Mental Health Clinic
```

SOR - Sundhedsvæsenets Organisationsregister - Danish Healthcare Organization Register

Example

A clinic recorded under a doctor's name

clinic record

```
Name: Michael Skjelbo Nielsen
Type: supplerende oplysninger
Part of: Hjerteklinik Nordjylland I/S
Inst. type: speciallægepraksis
Speciality: intern medicin
```

→

what came back

Ambulatory Medical Specialty Clinic / Center

`concept_id 38004240`

and its reason

"part of a speciallægepraksis with a specialty in intern medicin, which indicates outpatient specialty care without inpatient services"

The name was not helpful, so the LLM used the other fields

- **The fourth field contained the important information: speciallægepraksis → Ambulatory Medical Specialty Clinic**
- The LLM also provides a reason for final human review

PART 03

Interoperability through Inference and Statistical Trust

Paris Carbone

Quiz: How many interoperability conflicts do you see ?

Equipment Makers

Insurers & Payers

Hospital management

Regional Authorities

EU & National Legislators

Individual Clinicians

Quiz: How many interoperability conflicts do you see ?

Equipment Makers

Insurers & Payers

Hospital management

$$\binom{6+1}{2} = \frac{6(6+1)}{2} = 21$$

Regional Authorities

EU & National Legislators

Individual Clinicians

A fragmentation issue

- HL7, SNOMED CT, LOINC, ICD each solved their own layer, on their own timeline, at a technical level
- ICD (International Classification of Diseases) has met the most success…as a cash register (billing, insurance)
- **organizational interoperability remains elusive because hospitals adopting the same standards still produce semantically incompatible datasets** (as noted by Healthcare Information and Management Systems Society - HIMSS)

Can one standard save the day?

	HOSPITAL A	HOSPITAL B
blood-pressure episode	**1 row** [start date, end date], spanning three days	**3 rows** one per day, same clinical event
inflammation marker	**called CRP**	**called PCR**

PCR stands for **polymerase chain reaction**, while in Spanish, **PCR is C-reactive protein.**

! **We can agree on the envelop (e.g., HL7). We need to be able to map the contents**

From Transformation to Inference

GEN 2

Map every source into one rigid schema

What does not fit is dropped

The loss is structural

GEN 3

Decouple representation from any schema

Align subgraphs

Infer the missing links between concepts

de Mello 2022 · Abad-Navarro 2024 · Zhou et al. 2022 (cross-institutional EHR KG integration) · FHIR Lens, CBMS 2025 · CogKR · BioHopR, ACL 2025 · Cappuzzo, EDBT 2024

Pick your “Mapping” flavor

EQUIVALENCE · A ≡ B

These two mean the same thing

NCIT Gastric Carcinoma
≡ DOID stomach cancer

Most researched

SUBSUMPTION · A ⊑ B

This is a kind of that

KTPJ05E ⊑ SNOMED lymph node biopsy

More common, less researched.

NORMALISATION text → concept

Attach free-text to a standard concept

DE119 → SNOMED 201826

The entry point for real EHR data.

All three mappings can be modeled as a link prediction problem in a graph.

	Candidate Pairs
NCIT–DOID	133 million
SNOMED–FMA (body)	3.1 billion
SNOMED × FMA	37 billion
SNOMED × NCIT	50 billion

Anatomy of Modern Ontology Mapping

CANDIDATE GENERATION
cheap, high recall
narrow millions of pairs to hundreds

›

VERIFICATION
expensive, high precision
score each surviving pair

›

GLOBAL FILTER
cardinality, threshold, coherence
turn scores into an alignment

Lexical	AgreementMakerLight, LogMapLt	candidate generation
Logic + repair	LogMap, LogMapBio	global filter, the only family with a coherence check
Language models	BERTMap, BERTSubs, SapBERT	verification
Ontology / graph embeddings	OWL2Vec*, OWL2Vec4OA, BioGITOM	generation + verification
KG entity alignment	GCN-Align, BootEA, RDGCN	verification
LLM-based	OLaLa, LLMs4OM, Agent-OM, LogMap-LLM	verification
Hybrid / ensemble	Matcha, HybridOM	all three

“

When a measure becomes a target, it ceases to be a good measure

Charles Goodhart's Law

Let’s talk about Efficacy

Ontology Alignment Evaluation Initiative’s **Bio-ML benchmark** : The “TPC” of medical interoperability since 2022

① GLOBAL MATCHING

Output a complete alignment → P / R / F1

“Can you produce an alignment someone could actually use?”

② LOCAL RANKING

Rank the truth against 100 negatives → MRR, Hits@1

“Can you tell the right answer from a hundred wrong ones?”

The reference mappings are known to be incomplete. Under global matching, a system that proposes a **correct** mapping that is missing from the gold standard is punished as a false positive. Fixing the candidate set removes that penalty.

Focuses on matching whole biomedical ontologies

OMIM – ORDO	NCIT – DOID	SNOMED – FMA	SNOMED – NCIT	SNOMED – NCIT
disease	disease	body	pharm	neoplasm

Equivalence and subsumption · unsupervised and semi-supervised · references derived from Mondo and from UMLS concept identifiers

Ranking is solved. Alignment not yet

0.89 – 0.97

best MRR, local ranking
every task, every year

essentially saturated

0.72 – 0.92

best F1, global matching
across the five pairs

and flat since 2023

0.33 – 0.70

best Hits@1, subsumption
ranking

one genuine entrant, ever

GLOBAL-MATCHING F1, PAIR BY PAIR

Contesters differ each year

	2023	2024	2025
OMIM – ORDO	0.715	0.715	0.715
NCIT – DOID	0.913	0.918	0.908
SNOMED – FMA	0.790	0.790	0.790
SNOMED – NCIT (pharm)	0.790	0.902	0.760
SNOMED – NCIT (neoplasm)	0.771	0.771	0.782

OAEI Bio-ML results, 2023–2025 (CEUR-WS).

LogMapBio, logic-based, 2011 is still the best unsupervised result on two of the five pairs in 2025.

Where the Difficulty Actually Lives

A RECALL PROBLEM

OMIM – ORDO

LogMapLt: P = 0.94, R = 0.25. Even the best system reaches only R = 0.61.

Gene- and phenotype-style names share almost no surface form with Orphanet labels.

A THRESHOLDING PROBLEM

SNOMED – NCIT (neoplasm)

BERTMap P = 0.56 / R = 0.76. LogMapLt P = 0.95 / R = 0.52. Opposite operating points, near-identical F1.

Compositional terms
site + morphology + behaviour.

THE EASY ONE

NCIT – DOID

Best F1 above 0.90, and subsumption is near-solved here too.

Two disease ontologies with heavily overlapping English labels.

! **Only the first is a problem of representation. The second is a problem of deciding where to cut.**
Two systems can agree on every score and still disagree completely about which mappings to commit.

Does the Structure Earn Its Place?

ABLATIONS

BERTMap · AAAI 2022, SNOMED–FMA

Prediction only 0.832 → + structure 0.832 → + repair 0.836. The text-corpus ablation moves F1 from 0.350 to 0.832.

LogMap-ML · ESWC 2021

Adding path context made it worse: 0.885 with paths against 0.927 without.

LLMs4OM · ESWC 2024

Parent/child context helped in some domains and degraded performance on the Bio-ML tasks.

RE-EVALUATIONS, FROM the DB community

OpenEA · PVLDB 2020

Embedding-based entity alignment showed “no superiority over the conventional approaches PARIS and LogMap”.

Leone et al. · PVLDB 2022

significantly outperforms all representative neural methods; ~80% of the gap is entities matchable by simple string matching.

OM as link prediction · OM 2025

17 KGE models over the merged ontologies: 74.9% F1 on a structure-rich, label-poor task, but 31.8% vs BERTMap's 64.6% on OMIM–ORDO.

Structure earns its place in domains that are structure-rich and label-poor.

Biomedical vocabularies are label-rich. Which makes structural methods struggle to justify themselves here.

What If the Wrong Line Is Drawn?

The failure is silent, the shortcut is invisible, and both move when the data moves

The model may be learning the wrong regularity

In biomedical knowledge-graph completion, accuracy tracks how well-connected a node is. Well-connected concepts match well-connected concepts whether or not they mean the same thing.

And it moves when you move site

Coding practice, episode structure and graph topology all differ across hospitals. A system tuned on one site is not tuned on the next.

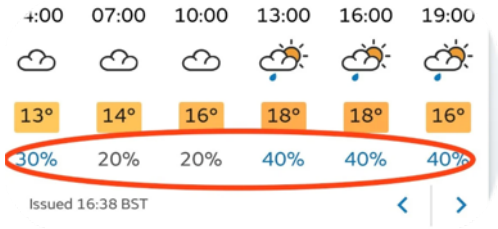
WHAT WE ACTUALLY WANT

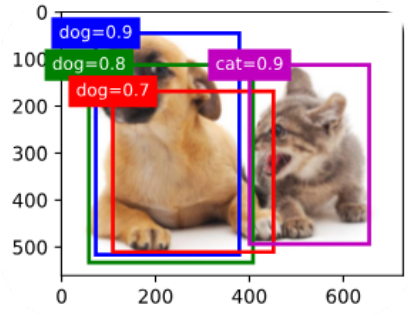
“I am not sure about this one. Can you take a look?”

A system that knows which of its own outputs to escalate based on the uncertainty.

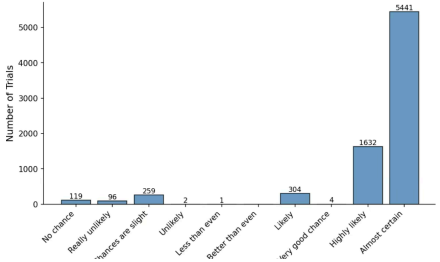
! **The system already produces a confidence score. It just isn't a probability.**
A 0.9 does not mean a nine-in-ten chance of being right and nobody knows what it does mean.

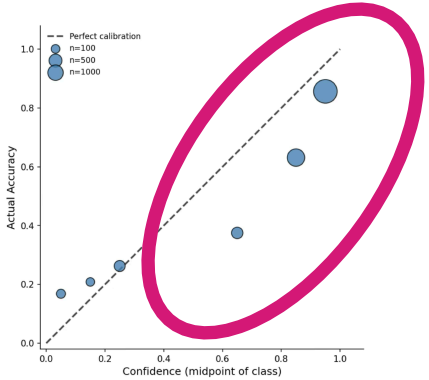
Uncertainty for decision processes





Asking our models to give us confidence

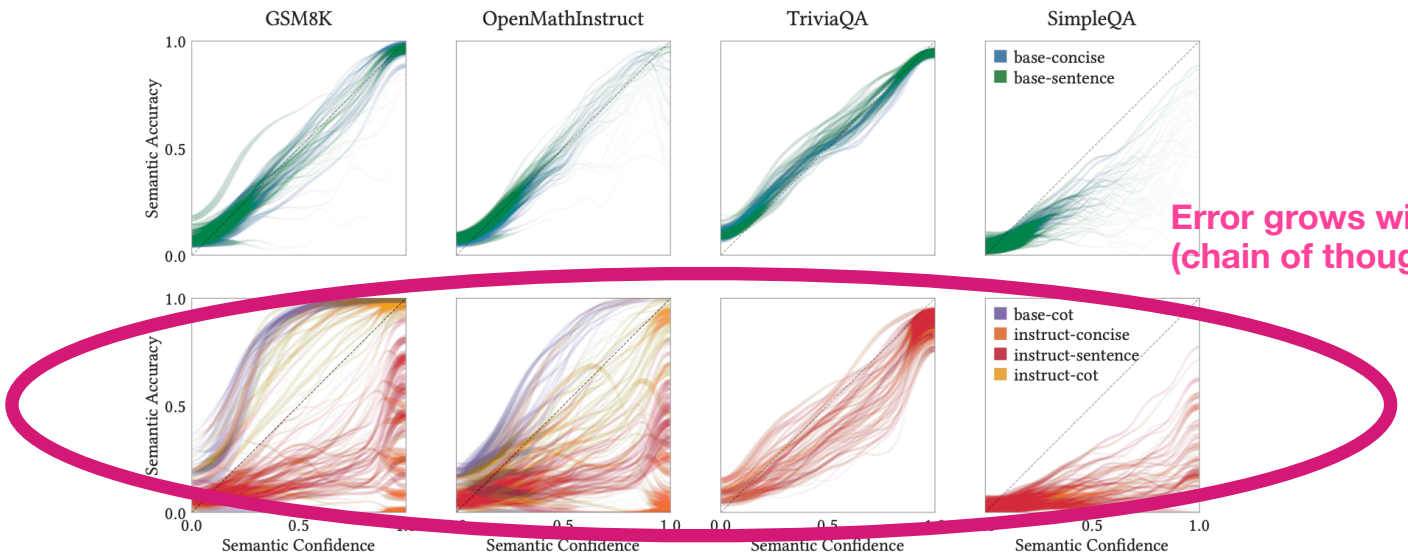




Overconfident zone

How do LLMs Compute Verbal Confidence? - Kumaran et al @Deepmind arXiv March 2026

Where calibration breaks



Error grows with capabilities
(chain of thought, reasoning RL)

Trained on Tokens, Calibrated on Concepts: The Emergence of Semantic Calibration in LLMs - Nakkiran et al @Apple arXiv Nov 6 2025

Conformal Prediction for calibration

High Confidence — Low Confidence

{ fox squirrel 0.99 }

{ fox squirrel 0.82, gray fox 0.03, bucket 0.02, rain barrel 0.02 }

{ marmot 0.30, fox squirrel 0.22, mink 0.18, weasel 0.16, beaver 0.03, polecat 0.01 }

Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal Prediction: A Gentle Introduction. Found. Trends Mach. Learn. 16, 4 (Mar 2023), 494–591. https://doi.org/10.1561/2200000101

Building a Safety Net

Model-agnostic
A GNN, an LLM, or a string matcher is treated as a black box, and never retrained.

Distribution-free
No assumption that the data is Gaussian, or i.i.d., or that the model is any good.

Finite-sample
A few hundred examples where you know the answer is enough to get started.

From a single answer → to a **set** that is guaranteed to contain the truth at a rate **you** choose. The worse the model/harder the mapping, the larger the set.

AND BEYOND COVERAGE — CONFORMAL RISK CONTROL

Bound any risk you can write down. Tumour segmentation: cap the false negatives, and as uncertainty rises the predicted region widens, until it is too wide to act on, and the case goes to a radiologist.

PART 03

Interactive Data Harmonization

Christos Koutras

Interactive Data Harmonisation

From AI-suggested mappings to decisions experts can trust

- LLMs facilitate semantic alignment, but a mapping is a decision not a prediction – errors matter
- Suggestions must be inspected, validated, and refined by domain experts before they can be trusted.
- How to build tools where AI does the heavy lifting, but the experts stay in control?

Interactive Data Harmonisation

From AI-suggested mappings to decisions experts can trust

- LLMs facilitate semantic alignment, but a mapping is a decision not a prediction – errors matter
- Suggestions must be inspected, validated, and refined by domain experts before they can be trusted.
- How to build tools where AI does the heavy lifting, but the experts stay in control?

AI proposes
matchers & LLMs generate candidate mappings, at scale

Interactive Data Harmonisation

From AI-suggested mappings to decisions experts can trust

- LLMs facilitate semantic alignment, but a mapping is a decision not a prediction – errors matter
- Suggestions must be inspected, validated, and refined by domain experts before they can be trusted.
- How to build tools where AI does the heavy lifting, but the experts stay in control?

AI proposes
matchers & LLMs generate candidate mappings, at scale

The expert stays in control
inspect · validate · refine with evidence

Interactive Data Harmonisation

From AI-suggested mappings to decisions experts can trust

- LLMs facilitate semantic alignment, but a mapping is a decision not a prediction – errors matter
- Suggestions must be inspected, validated, and refined by domain experts before they can be trusted.
- How to build tools where AI does the heavy lifting, but the experts stay in control?

AI proposes
matchers & LLMs generate candidate mappings, at scale

The expert stays in control
inspect · validate · refine with evidence

A result you can trust
correct, reproducible mappings

Interactive Data Harmonisation

From AI-suggested mappings to decisions experts can trust

- LLMs facilitate semantic alignment, but a mapping is a decision not a prediction – errors matter
- Suggestions must be inspected, validated, and refined by domain experts before they can be trusted.
- How to build tools where AI does the heavy lifting, but the experts stay in control?

*AI-based tools should handle the bulk of alignment and disambiguation - **practitioners should retain control over the final results***

AI proposes
matchers & LLMs generate candidate mappings, at scale

The expert stays in control
inspect · validate · refine with evidence

A result you can trust
correct, reproducible mappings

What does Harmonisation mean?

Running Example: Pooling cancer cohorts into the NCI Genomics Data Commons (GDC)

10 cancer cohorts — one shown

Age	Histologic_Grade_FIGO	…
64	FIGO grade 2	…
71	FIGO grade 3	…

179–213 columns each, own naming & units

One target: the Genomic Data Commons (GDC)

age_at_diagnosi	tumor_grade	…
23,376	G2	…
25,933	G3	…

700+ attributes with controlled values

What does Harmonisation mean?

Running Example: Pooling cancer cohorts into the NCI Genomics Data Commons (GDC)

10 cancer cohorts — one shown

Age	Histologic_Grade_FIGO	…
64	FIGO grade 2	…
71	FIGO grade 3	…

179–213 columns each, own naming & units

One target: the Genomic Data Commons (GDC)

age_at_diagnosi	tumor_grade	…
23,376	G2	…
25,933	G3	…

700+ attributes with controlled values

1 Schema Matching

`Histologic_Grade_FIGO → tumor_grade`

What does Harmonisation mean?

Running Example: Pooling cancer cohorts into the NCI Genomics Data Commons (GDC)

10 cancer cohorts — one shown

Age	Histologic_Grade_FIGO	…
64	FIGO grade 2	…
71	FIGO grade 3	…

179–213 columns each, own naming & units

One target: the Genomic Data Commons (GDC)

age_at_diagnosi	tumor_grade	…
23,376	G2	…
25,933	G3	…

700+ attributes with controlled values

1 Schema Matching

`Histologic_Grade_FIGO → tumor_grade`

2 Value Matching

`FIGO grade 2 → G2`

What does Harmonisation mean?

Running Example: Pooling cancer cohorts into the NCI Genomics Data Commons (GDC)

10 cancer cohorts — one shown

Age	Histologic_Grade_FIGO	…
64	FIGO grade 2	…
71	FIGO grade 3	…

179–213 columns each, own naming & units

One target: the Genomic Data Commons (GDC)

age_at_diagnosi	tumor_grade	…
23,376	G2	…
25,933	G3	…

700+ attributes with controlled values

1 Schema Matching

`Histologic_Grade_FIGO → tumor_grade`

2 Value Matching

`FIGO grade 2 → G2`

3 Transformation

`Age 64 (years) → 23,376 (days) ×365.25`

Complexity of Schemata

One source column… **…three plausible GDC targets. Which one?**

Histologic_Grade_FIGO
FIGO grade 1
FIGO grade 2
FIGO grade 3
…

?

tumor_grade	who_nte_grade	adverse_event_grade
G1	G1	Grade 1
G2	G2	Grade 2
G3	G3	Grade 3
…	…	…

! Each cohort has 179–213 of them, against 700+ GDC attributes: numerous candidate pairs.

Complexity of Values

Age
28
59
76

?

age_at_diagnosis
18250
29220
21915

Age at diagnosis, in days since birth.

age_at_onset
55
72
91

Age when exposure to a specific environmental factor began.

age_at_index
60
75
58

Age in years at the anchor date used for de-identification.

! Same concept, three encodings — only definitions and values can disambiguate.

Why Automation Isn't Enough

Nuanced semantics and scale of verification

NUANCED SEMANTICS

age_at_diagnosis
days since birth at diagnosis — e.g. 18250, 29220

age_at_onset
age (in years) when exposure to a specific factor began

age_at_index
age (in years) at the anchor date used for de-identification

Only one correct match for **Age** target attribute

SCALE OF VERIFICATION

- Even the best automated matcher (Magneto, VLDB '25) has Recall < 0.5 on GDC mapping
- To reach acceptable recall, experts must review top-k lists with a *small* k
- …that is 1000+ candidate matches per dataset!
- Today this happens in Excel, R, and unpublished one-off scripts – slow, error-prone, irreproducible

Why Automation Isn't Enough

Nuanced semantics and scale of verification

NUANCED SEMANTICS

age_at_diagnosis
days since birth at diagnosis — e.g. 18250, 29220

age_at_onset
age (in years) when exposure to a specific factor began

age_at_index
age (in years) at the anchor date used for de-identification

Only one correct match for **Age** target attribute

SCALE OF VERIFICATION

- Even the best automated matcher (Magneto, VLDB '25) has Recall < 0.5 on GDC mapping
- To reach acceptable recall, experts must review top-k lists with a *small* k
- …that is 1000+ candidate matches per dataset!
- Today this happens in Excel, R, and unpublished one-off scripts – slow, error-prone, irreproducible

The bottleneck is not generating candidate mappings — it is verifying them.
Interactive tools must optimize expert attention, not just model accuracy.

Three Ways to Keep the Expert in Control

Three interaction surfaces, chosen by user and workflow

Three Ways to Keep the Expert in Control

Three interaction surfaces, chosen by user and workflow

PARADIGM ①

Harmonization as code

Composable primitives inside notebooks and pipelines — scriptable, versionable, reproducible.

for data scientists & pipeline builders

Three Ways to Keep the Expert in Control

Three interaction surfaces, chosen by user and workflow

PARADIGM ①

Harmonization as code

Composable primitives inside notebooks and pipelines — scriptable, versionable, reproducible.

for data scientists & pipeline builders

PARADIGM ②

Harmonization as visual curation

The whole candidate space on one screen: overview first, evidence on demand, direct accept / reject.

for domain experts curating at scale

Three Ways to Keep the Expert in Control

Three interaction surfaces, chosen by user and workflow

PARADIGM ①

Harmonization as code

Composable primitives inside notebooks and pipelines — scriptable, versionable, reproducible.

for data scientists & pipeline builders

PARADIGM ②

Harmonization as visual curation

The whole candidate space on one screen: overview first, evidence on demand, direct accept / reject.

for domain experts curating at scale

PARADIGM ③

Harmonization as dialogue

Agents plan, call the same primitives, self-check — and ask the human when unsure.

for practitioners who don't program

Three Ways to Keep the Expert in Control

Three interaction surfaces, chosen by user and workflow

PARADIGM ①

Harmonization as code

Composable primitives inside notebooks and pipelines — scriptable, versionable, reproducible.

for data scientists & pipeline builders

PARADIGM ②

Harmonization as visual curation

The whole candidate space on one screen: overview first, evidence on demand, direct accept / reject.

for domain experts curating at scale

PARADIGM ③

Harmonization as dialogue

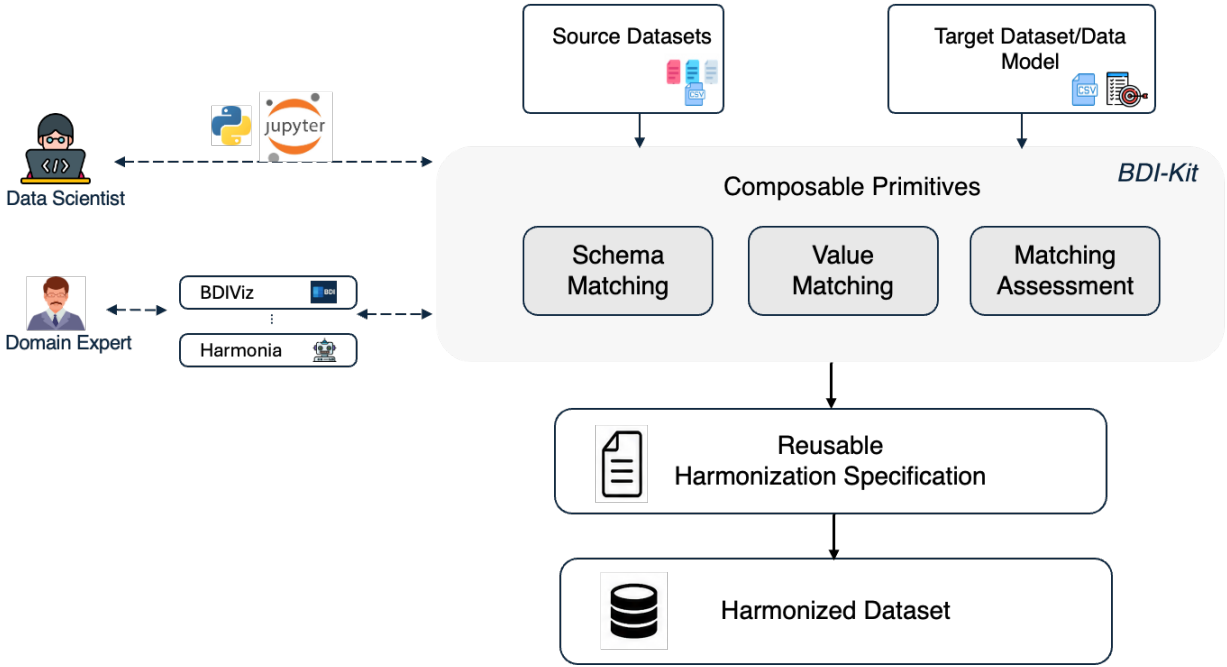
Agents plan, call the same primitives, self-check — and ask the human when unsure.

for practitioners who don't program

IN THIS SEGMENT
Three solutions built on one shared backbone

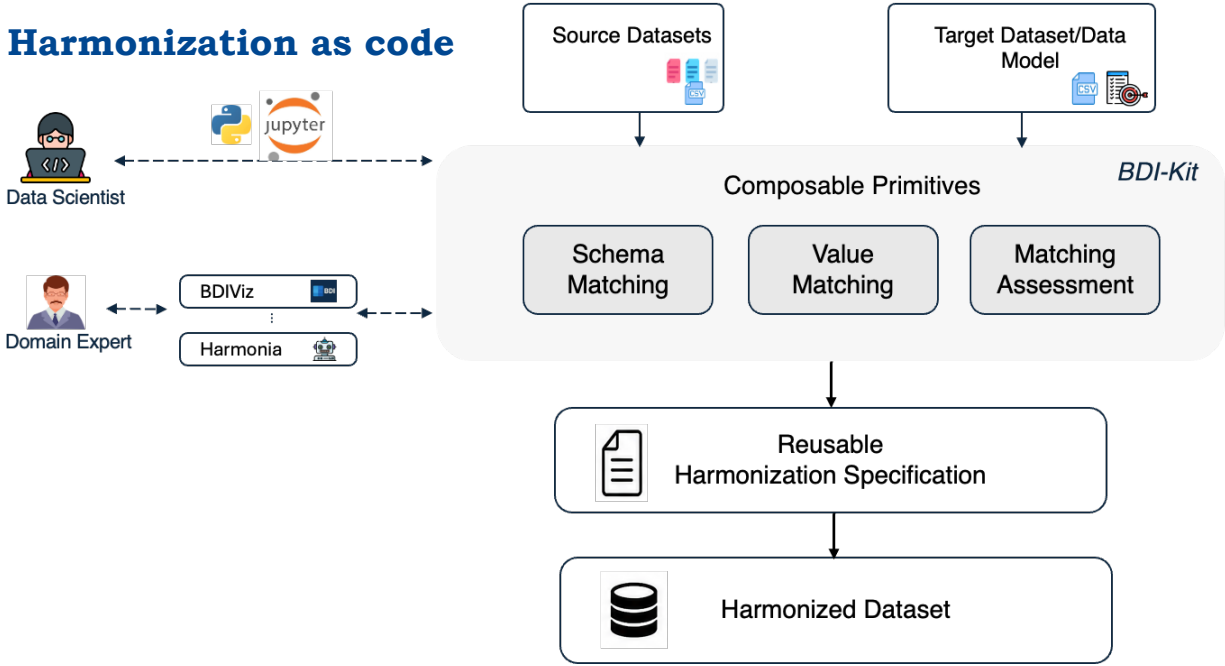
1 **BDIKit** 2 **BDIViz** 3 **Harmonia**

Shared Harmonization Backbone

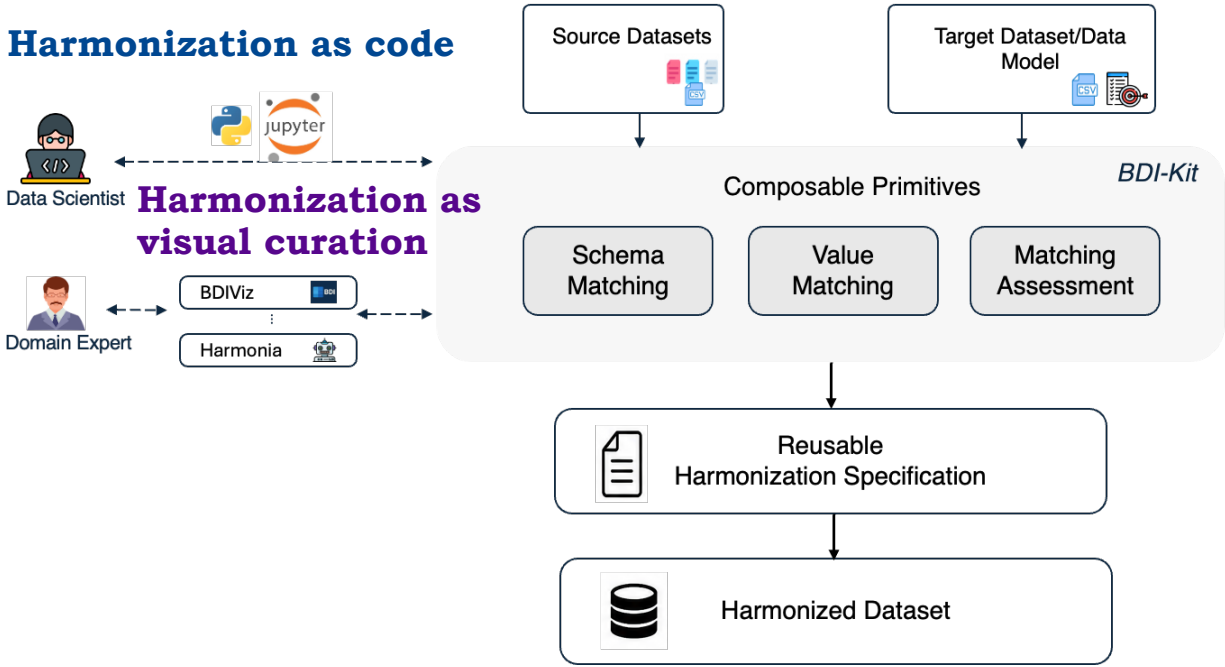


Shared Harmonization Backbone

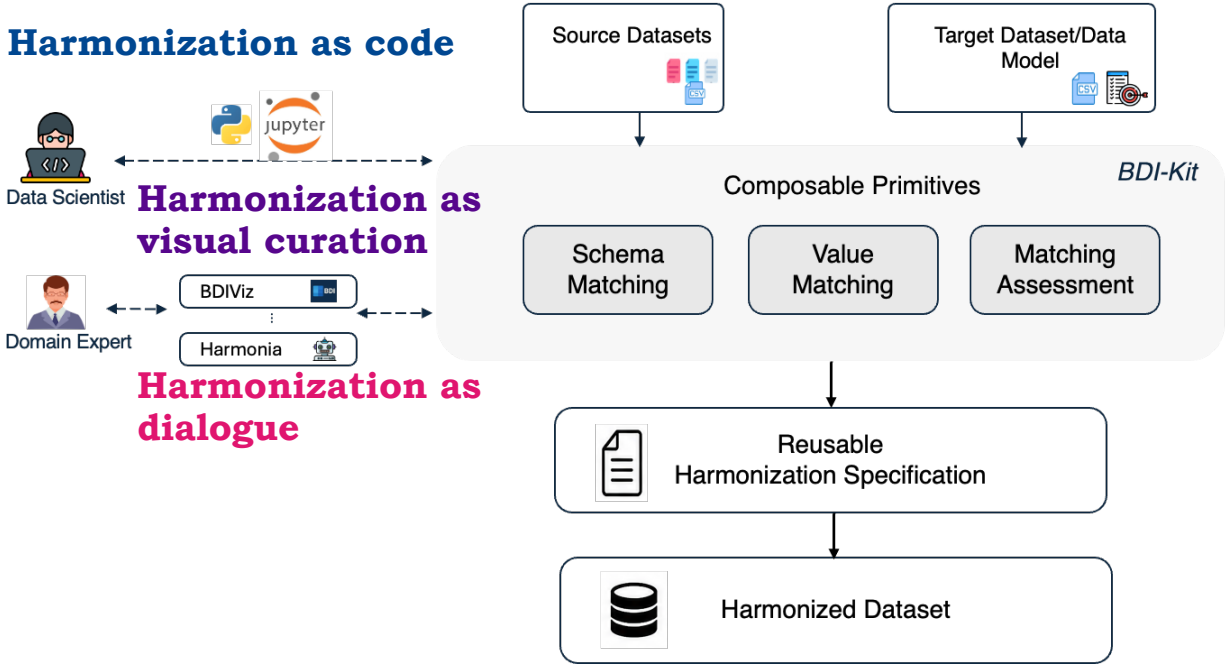
Harmonization as code



Shared Harmonization Backbone



Shared Harmonization Backbone



BDIKit – Harmonization as Code

Open-source primitives that make each sub-task a one-liner, and the whole workflow a script

```
import bdikit as bdi

# 1 — schema matching: cohort → GDC
m = bdi.match_schema(df, target="gdc",
        method="magneto_ft")
# …expert inspects & repairs a match
bdi.top_matches(df, "Histologic_type",
        target="gdc")

# 2 — value matching per attribute pair
v = bdi.match_values(df, target="gdc",
        attribute_matches=m, method="tfidf")

# 3 — harmonized table + reusable spec
gdc_df = bdi.materialize_mapping(df, v)
```

Primitives, not pipelines

match_schema · top_matches · match_values · materialize_mapping — compose with pandas.

Agent-ready via MCP

the same primitives exposed as MCP tools — any AI agent can drive bdi-kit in a chat, model-agnostically.

One API, many matchers

Magneto (SLM+LLM), contrastive learning, string & embedding similarity — swap one argument.

Reproducible by design

declarative JSON mapping specs — rerun on new cohorts, publish with the paper.

Standards built in

target="gdc" loads attribute names, definitions, and permissible values.

R. Lopez, A. Santos, C. Koutras, J. Freire — "BDI-Kit", Patterns (2025) · demo with MCP agents: SIGMOD '26 · github.com/VIDA-NYU/bdi-kit

BDIKit – Harmonization as Code

Open-source primitives that make each sub-task a one-liner, and the whole workflow a script

```
import bdikit as bdi

# 1 — schema matching: cohort → GDC
m = bdi.match_schema(df, target="gdc",
        method="magneto_ft")
# …expert inspects & repairs a match
bdi.top_matches(df, "Histologic_type",
        target="gdc")

# 2 — value matching per attribute pair
v = bdi.match_values(df, target="gdc",
        attribute_matches=m, method="tfidf")

# 3 — harmonized table + reusable spec
gdc_df = bdi.materialize_mapping(df, v)
```

Primitives, not pipelines

match_schema · top_matches · match_values · materialize_mapping — compose with pandas.

Agent-ready via MCP

the same primitives exposed as MCP tools — any AI agent can drive bdi-kit in a chat, model-agnostically.

One API, many matchers

Magneto (SLM+LLM), contrastive learning, string & embedding similarity — swap one argument.

Reproducible by design

declarative JSON mapping specs — rerun on new cohorts, publish with the paper.

Standards built in

target="gdc" loads attribute names, definitions, and permissible values.

Verification still means inspecting DataFrames — for 700+ attributes we can do better →

R. Lopez, A. Santos, C. Koutras, J. Freire — "BDI-Kit", Patterns (2025) · demo with MCP agents: SIGMOD '26 · github.com/VIDA-NYU/bdi-kit

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE

Too many candidate pairs

700+ target attributes, deep top-k lists → 1,000+ pairs to review

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

→

DC1 · Efficient discovery
Surface likely matches automatically — eliminate tedious one-by-one inspection.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

DC1 · Efficient discovery
Surface likely matches automatically — eliminate tedious one-by-one inspection.

DC2 · Interactive filtering
Slice the candidate space by ontology group, datatype, and confidence.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

CHALLENGE
Nuanced semantics
names mislead — multiple seemingly equivalent candidates

DC1 · Efficient discovery
Surface likely matches automatically — eliminate tedious one-by-one inspection.

DC2 · Interactive filtering
Slice the candidate space by ontology group, datatype, and confidence.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

DC1 · Efficient discovery
Surface likely matches automatically — eliminate tedious one-by-one inspection.

DC2 · Interactive filtering
Slice the candidate space by ontology group, datatype, and confidence.

CHALLENGE
Nuanced semantics
names mislead — multiple seemingly equivalent candidates

DC3 · Value-based evidence
Show value distributions and comparisons, not just attribute-name similarity.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

CHALLENGE
Nuanced semantics
names mislead — multiple seemingly equivalent candidates

DC1 · Efficient discovery
Surface likely matches automatically — eliminate tedious one-by-one inspection.

DC2 · Interactive filtering
Slice the candidate space by ontology group, datatype, and confidence.

DC3 · Value-based evidence
Show value distributions and comparisons, not just attribute-name similarity.

DC4 · Agent validation & feedback
LLM explanations for candidate matches — validated by the user, folded back into curation.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

CHALLENGE
Nuanced semantics
names mislead — multiple seemingly equivalent candidates

CHALLENGE
Curation is a process
multi-session, multi-source work — mistakes must be reversible

DC1 · Efficient discovery
Surface likely matches automatically — eliminate tedious one-by-one inspection.

DC2 · Interactive filtering
Slice the candidate space by ontology group, datatype, and confidence.

DC3 · Value-based evidence
Show value distributions and comparisons, not just attribute-name similarity.

DC4 · Agent validation & feedback
LLM explanations for candidate matches — validated by the user, folded back into curation.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

→ **DC1 · Efficient discovery**
Surface likely matches automatically — eliminate tedious one-by-one inspection.

DC2 · Interactive filtering
Slice the candidate space by ontology group, datatype, and confidence.

CHALLENGE
Nuanced semantics
names mislead — multiple seemingly equivalent candidates

→ **DC3 · Value-based evidence**
Show value distributions and comparisons, not just attribute-name similarity.

DC4 · Agent validation & feedback
LLM explanations for candidate matches — validated by the user, folded back into curation.

CHALLENGE
Curation is a process
multi-session, multi-source work — mistakes must be reversible

→ **DC5 · Workflow support**
Undo/redo, action timeline, and side-by-side comparison across sources.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

BDIViz: Design Requirements

Each challenge of the task maps to design considerations (DCs) — elicited in interviews and co-design

CHALLENGE
Too many candidate pairs
700+ target attributes, deep top-k lists → 1,000+ pairs to review

DC1 · Efficient discovery
Surface likely matches automatically — eliminate tedious one-by-one inspection.

DC2 · Interactive filtering
Slice the candidate space by ontology group, datatype, and confidence.

CHALLENGE
Nuanced semantics
names mislead — multiple seemingly equivalent candidates

DC3 · Value-based evidence
Show value distributions and comparisons, not just attribute-name similarity.

DC4 · Agent validation & feedback
LLM explanations for candidate matches — validated by the user, folded back into curation.

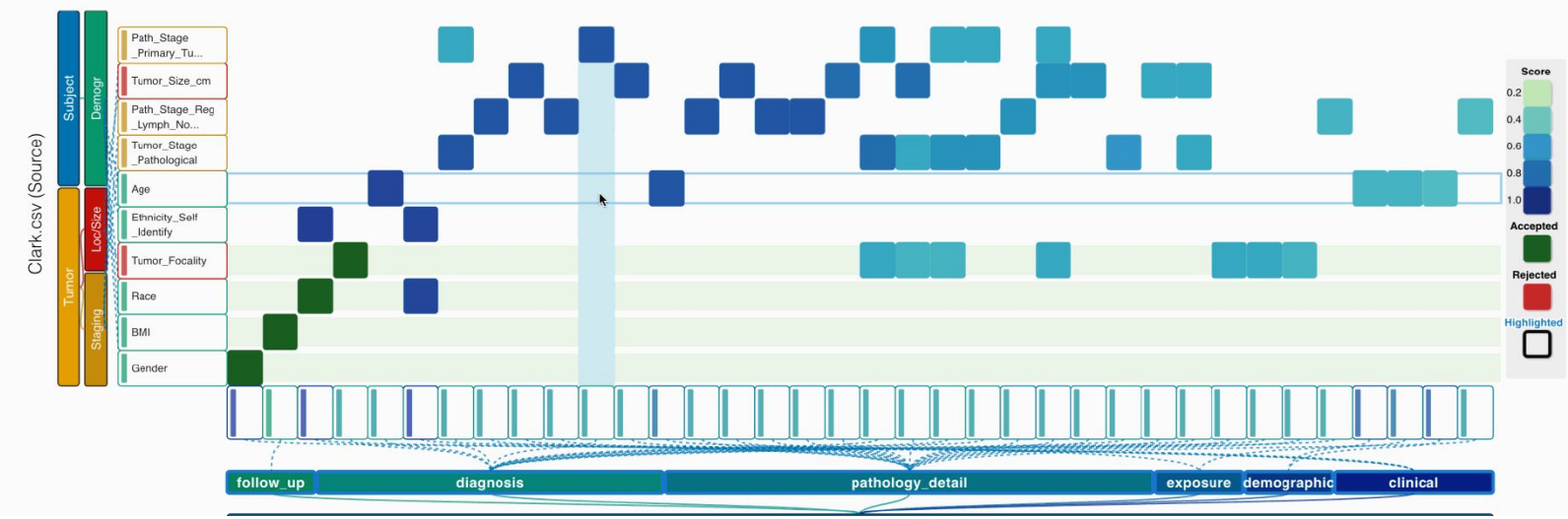
CHALLENGE
Curation is a process
multi-session, multi-source work — mistakes must be reversible

DC5 · Workflow support
Undo/redo, action timeline, and side-by-side comparison across sources.

Design lesson: requirements were elicited from failing workflows — then iterated in co-design.

E. Wu et al. — "BDIViz: Interactive Visualization for Biomedical Schema Matching with LLM-Powered Validation", IEEE TVCG / VIS 2025

Candidate Heatmap



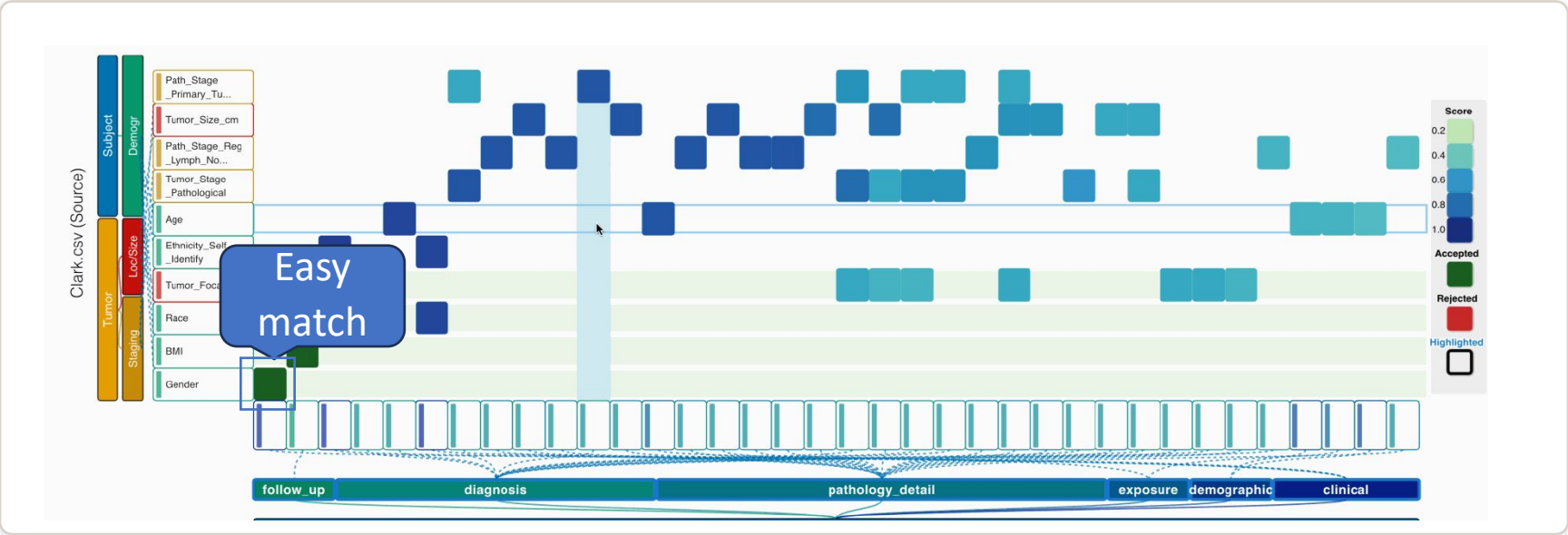
pool many matchers → recall ↑

color = ensemble score → triage hard cases

hierarchical axes: ontology groups of 700+ attrs

green accepted · red rejected

Candidate Heatmap



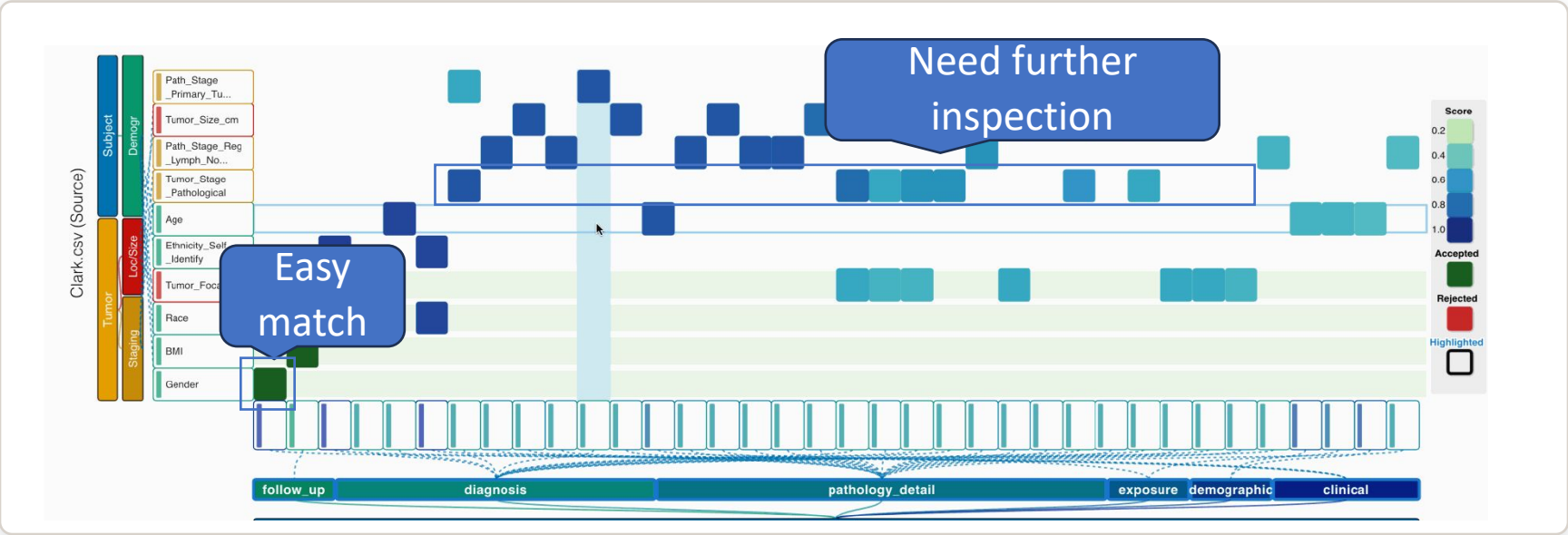
pool many matchers → recall ↑

color = ensemble score → triage hard cases

hierarchical axes: ontology groups of 700+ attrs

green accepted · red rejected

Candidate Heatmap



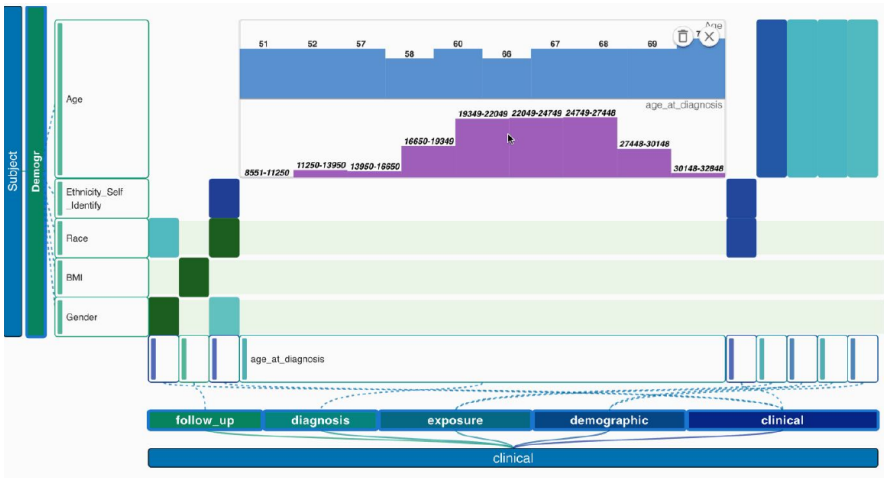
pool many matchers → recall ↑

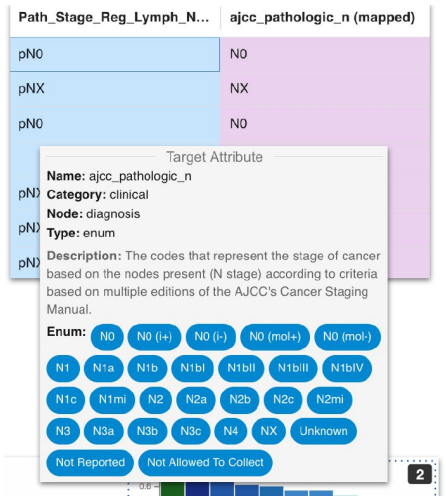
color = ensemble score → triage hard cases

hierarchical axes: ontology groups of 700+ attrs

green accepted · red rejected

Per Candidate Evidence





value mapping + target definition, ontology node & permissible values

compare value distributions & samples

read definitions and enum domains

accept · reject · edit — inline

every action undoable, on a timeline

Age 51–69 (years) vs **age_at_diagnosis** 8,551–32,846 (days)

Validating Candidates with LLM Explanations

Four independent arguments per candidate

Value-based · semantic · knowledge-based · pattern — each with its own confidence score.

Grounded in the data

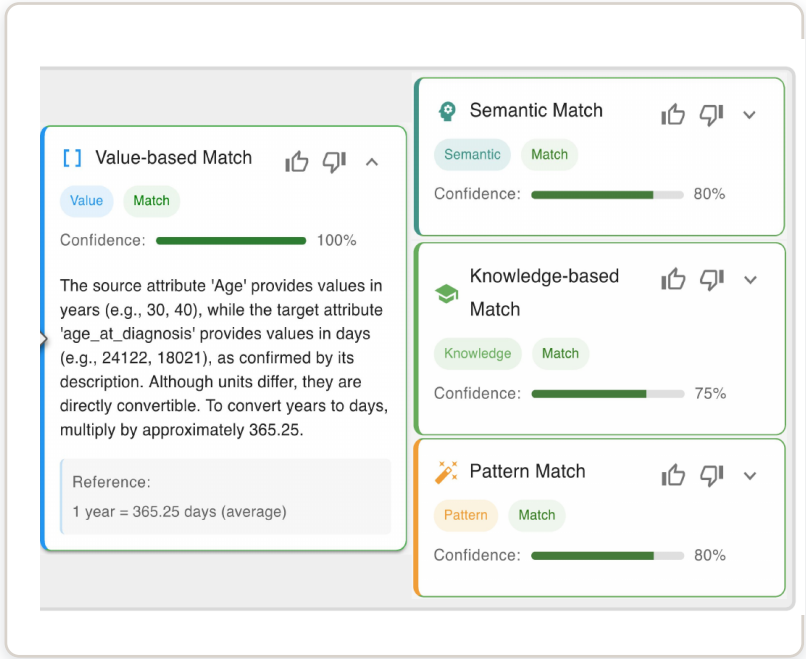
Explanations cite values and definitions — e.g. years vs. days, convert ×365.25.

Feedback becomes memory

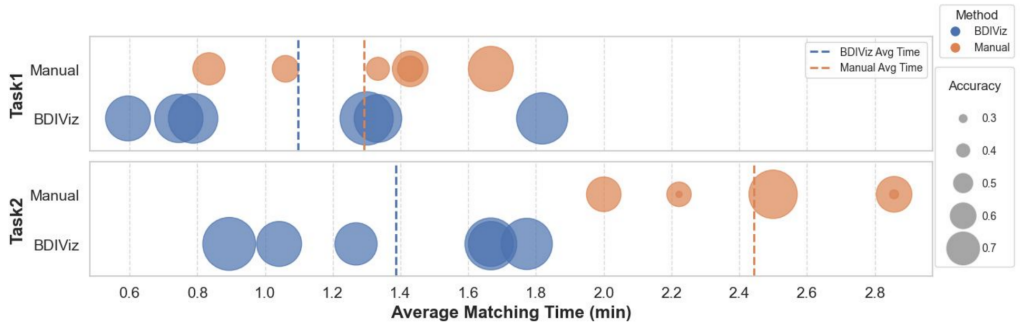
Thumbs-up / thumbs-down and past operations are retrieved to refine later reasoning.

The human still decides

The agent explains and flags — it never commits a mapping on its own.



Evaluation with 12 Domain Experts



Task 1: kidney cohort → GDC (top) · Task 2: lung cohort → PDC (bottom) · bubble size = accuracy

↑ accuracy

largest gains on abstract names: cT, cM, GN, OS

≈ −40% time

avg matching time, Task 2: 2.4 → 1.4 min

SUS 71.9

strong perceived usability across participants

TLX ↓

lower mental demand, effort & frustration

Study tasks from Li et al., Cell Rep. Med. 2024 (kidney) and Woldmar et al., ESMO Open 2023 (lung). BDIViz: Wu et al., IEEE TVCG / VIS 2025.

Harmonization as Dialogue

Route ① BDIKit's MCP tools — chat from any agent client.

Route ② Harmonia — adds the agent loop, notebook UI, provenance.

Expert
chat / notebook

↔

LLM agent
plan · act · check
= **Harmonia**/**any MCP client**

↔

BDIKit primitives

generated code

→

harmonized table
+ declarative spec

provenance DB — every step logged

Santos et al. — "Interactive Data Harmonization with LLM Agents", NOVAS @ SIGMOD 2025 · BDIKit MCP tools: Lopez et al., SIGMOD '26 demo

Harmonization as Dialogue

Route ① BDIKit's MCP tools — chat from any agent client.

Route ② Harmonia — adds the agent loop, notebook UI, provenance.

Expert
chat / notebook

↔

LLM agent
plan · act · check
= Harmonia/any MCP client

provenance DB — every step logged

↔

BDIKit primitives

generated code

→

harmonized table
+ declarative spec

example: a Harmonia session

Expert "Match this cohort to GDC using ct_learning."

Santos et al. — "Interactive Data Harmonization with LLM Agents", NOVAS @ SIGMOD 2025 · BDIKit MCP tools: Lopez et al., SIGMOD '26 demo

Harmonization as Dialogue

Route ① BDIKit's MCP tools — chat from any agent client.

Route ② Harmonia — adds the agent loop, notebook UI, provenance.

Expert
chat / notebook

↔

LLM agent
plan · act · check
= Harmonia/any MCP client

↔

BDIKit primitives

generated code

→

harmonized table
+ declarative spec

provenance DB — every step logged

example: a Harmonia session

Expert "Match this cohort to GDC using ct_learning."
Agent match_schema(target="gdc") → flags **Histologic_type → roots** as suspicious → top_matches() → proposes **primary_diagnosis** — accept? **Expert ✓**

Santos et al. — "Interactive Data Harmonization with LLM Agents", NOVAS @ SIGMOD 2025 · BDIKit MCP tools: Lopez et al., SIGMOD '26 demo

Harmonization as Dialogue

Route ① BDIKit's MCP tools — chat from any agent client.

Route ② Harmonia — adds the agent loop, notebook UI, provenance.

Expert
chat / notebook

↔

LLM agent
plan · act · check
= **Harmonia**/**any MCP client**

provenance DB — every step logged

↔

BDIKit primitives

generated code

→

harmonized table
+ declarative spec

example: a Harmonia session

Expert "Match this cohort to GDC using ct_learning."
Agent match_schema(target="gdc") → flags **Histologic_type → roots** as suspicious → top_matches() → proposes **primary_diagnosis** — accept? **Expert ✓**
Agent match_values() on 8 columns → catches **"II" → "Stage III"**, corrects to **"Stage II"** → materialize_mapping() → table + spec

Santos et al. — "Interactive Data Harmonization with LLM Agents", NOVAS @ SIGMOD 2025 · BDIKit MCP tools: Lopez et al., SIGMOD '26 demo

Design Lessons

Design Lessons

AI proposes, experts dispose

Every suggestion is scored; none is committed without a human decision.

Design Lessons

AI proposes, experts dispose

Every suggestion is scored; none is committed without a human decision.

Evidence, not just scores

Values, distributions, definitions, and explanations at the point of decision.

Design Lessons

AI proposes, experts dispose

Every suggestion is scored; none is committed without a human decision.

Evidence, not just scores

Values, distributions, definitions, and explanations at the point of decision.

Triage expert attention

Route the expert to hard cases — where matchers disagree or confidence is low.

Design Lessons

AI proposes, experts dispose

Every suggestion is scored; none is committed without a human decision.

Evidence, not just scores

Values, distributions, definitions, and explanations at the point of decision.

Triage expert attention

Route the expert to hard cases — where matchers disagree or confidence is low.

Every decision leaves a trace

Provenance + declarative specs → reproducible, reusable, publishable mappings.

Design Lessons

AI proposes, experts dispose

Every suggestion is scored; none is committed without a human decision.

Evidence, not just scores

Values, distributions, definitions, and explanations at the point of decision.

Triage expert attention

Route the expert to hard cases — where matchers disagree or confidence is low.

Every decision leaves a trace

Provenance + declarative specs → reproducible, reusable, publishable mappings.

One backbone, many surfaces

Code, visual, conversational — same primitives; meet users where they work.

PART 04

The New Frontier

Andra Ionescu, Paris Carbone, Christos Koutras, Katja Hose, Jayanth Raghotama

Open Problems

What is still missing between a score and a guarantee

1. **Nobody has done this for alignment**
   The graph, LLM and imaging literatures all have conformal methods. Ontology and schema matching have scores, top-k lists and ensemble ranks. and no coverage guarantee.

2. **Who produces the calibration set?**
   Coverage needs ground truth. 500 verified SKS→SNOMED mappings, at whose cost and can that set be shared when the underlying records cannot?

3. **Marginal coverage is not conditional coverage**
   90% overall is compatible with 40% on rare disease codes , exactly the codes that matter. Mondrian CP helps; full conditional coverage is provably impossible.

4. **Expert feedback breaks the assumption**
   The moment corrections re-enter the calibration set, the data stop being exchangeable, and the interactive designs that follow depend on that loop.

5. **Cohort Statistics as a goal**
   Prediction-powered inference (Science, 2023) gives valid confidence intervals from a small verified sample plus a large machine-labelled one which may matter more than getting every code right.

Open Problems in Interactive Harmonization

Open Problems in Interactive Harmonization

Benchmarks for interactive harmonization*

End-to-end evaluation of human+AI workflows — beyond isolated matching metrics.

*Santos, A. et. al (2025). GDC-SM: The GDC Schema Matching Benchmark (Version 1.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14963588

Open Problems in Interactive Harmonization

Benchmarks for interactive harmonization*

End-to-end evaluation of human+AI workflows — beyond isolated matching metrics.

Calibrated uncertainty

Expose confidence users can act on — it decides when the human is called.

*Santos, A. et. al (2025). GDC-SM: The GDC Schema Matching Benchmark (Version 1.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14963588

Open Problems in Interactive Harmonization

Benchmarks for interactive harmonization*

End-to-end evaluation of human+AI workflows — beyond isolated matching metrics.

Calibrated uncertainty

Expose confidence users can act on — it decides when the human is called.

Reproducibility vs. LLM brittleness

Same prompt, different run. Specs + provenance help — guarantees don't exist yet.

*Santos, A. et. al (2025). GDC-SM: The GDC Schema Matching Benchmark (Version 1.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14963588

Open Problems in Interactive Harmonization

Benchmarks for interactive harmonization*

End-to-end evaluation of human+AI workflows — beyond isolated matching metrics.

Calibrated uncertainty

Expose confidence users can act on — it decides when the human is called.

Reproducibility vs. LLM brittleness

Same prompt, different run. Specs + provenance help — guarantees don't exist yet.

Interaction design

NL is ambiguous: blend dialogue with structured, action-oriented interfaces.

*Santos, A. et. al (2025). GDC-SM: The GDC Schema Matching Benchmark (Version 1.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14963588

Open Problems in Interactive Harmonization

Benchmarks for interactive harmonization*

End-to-end evaluation of human+AI workflows — beyond isolated matching metrics.

Calibrated uncertainty

Expose confidence users can act on — it decides when the human is called.

Reproducibility vs. LLM brittleness

Same prompt, different run. Specs + provenance help — guarantees don't exist yet.

Interaction design

NL is ambiguous: blend dialogue with structured, action-oriented interfaces.

Provenance-aware agents

Learn user preferences from curation history; explain any output's lineage.

*Santos, A. et. al (2025). GDC-SM: The GDC Schema Matching Benchmark (Version 1.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14963588

Open Problems

1. **Provenance tracking**
   Mapping provenance, data provenance

2. **Federated, privacy preserving mapping**
   Multi-site clinical data

3. **Quantifying semantic loss**
   What is the value of the data that is not mapped?

Traditional Health Data Integration

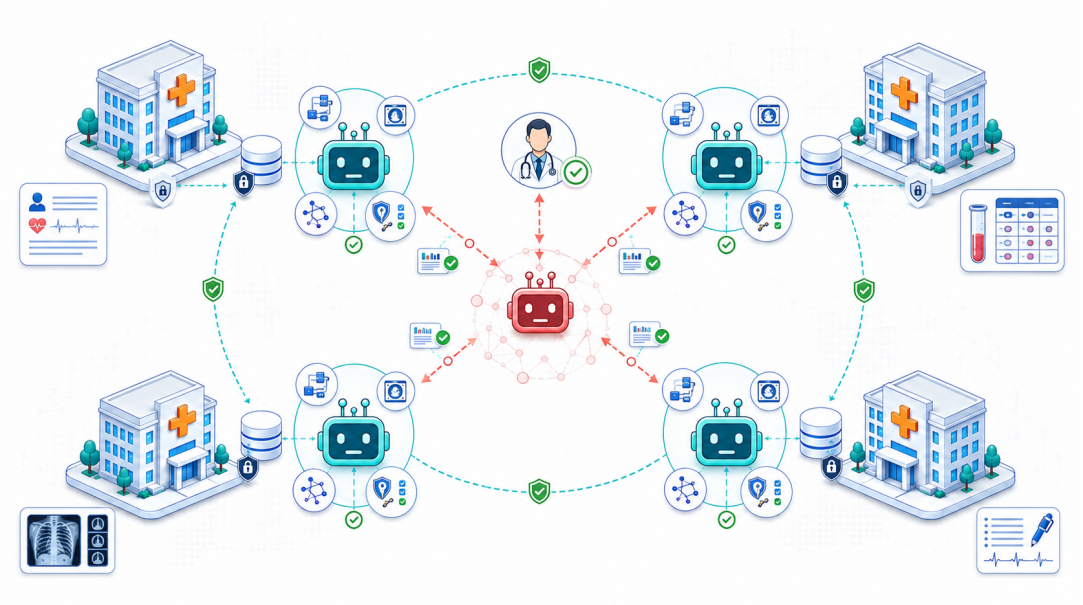
Central component

- Single point of failure
- Mapping data to a common schema
- Data harmonization and imputation
- Ontology-based data access
- Knowledge graphs

Agents in Healthcare

Central component

- Distributed system
- Collaboration and negotiation to achieve tasks
- Intelligently combine different skills and specialization
- Automatically align understanding
- Knowledge graphs
- Multi-modality



Conclusion

Why this matters

1. **Beyond Technology and Research**
   Improvement in data exchange can, and will lead to a tangible improvement in the delivery of care and outcomes for patients

2. **Data Quality**
   Better models, reduced data needs AND improvement in patient outcomes

3. **True Interoperability**
   Can only come from the domain agreeing on semantics

4. **But until then**
   Technology can fill the gaps, improve interoperability systems