

Symbolic AI for Medical Knowledge Representation & Reasoning: Just The Right Amount – SNOMED CT Content Extraction

Renate A. Schmidt

Formal Methods Research Group Leader
Department of Computer Science, University of Manchester

TRIPLE 26, Edinburgh, 29 March 2026

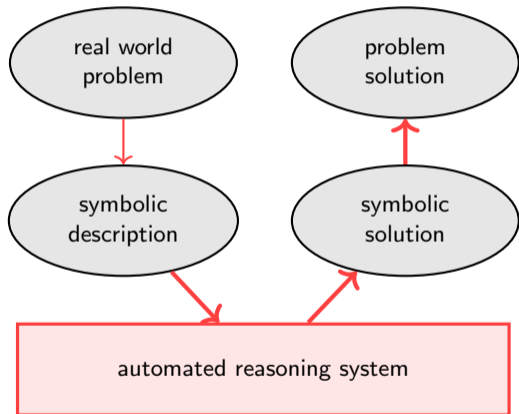
Symbolic AI for knowledge representation & reasoning

Gives us languages to describe information using **precise statements**, e.g., statements about computer programs, definitions of medical codes/concepts or other real world information

In addition gives us **logical techniques to reason** about these statements

Given a real world problem

1. Design a suitable symbolic description of the problem
2. Develop correct, practical automated reasoning system to find solution
3. Translate symbolic solution to problem solution



Automated reasoning systems

Provide powerful tools for

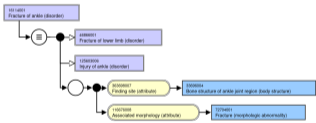
Ensuring correct functioning of complex software & hardware systems:

Safety critical applications, check/generate product configurations, ...
Major companies intensely using AR tools include Microsoft, Intel, AWS, NASA, Mercedes, Toyota, Airbus



Managing and retrieving data in complex information systems:

In medicine, underpinning services based on large knowledge bases of medical concepts, important for consistent creation, storage, auditing and analysis of patient health data



Many other difficult problems:

Performing combinatorial reasoning, professional sports scheduling, planning, optimisation, ...



Partnership with SNOMED Intl

Since 2019, we officially teamed up with SNOMED Intl, an international organisation distributing and maintaining the medical knowledge base SNOMED CT

Goal: Required an approach that produces concise extracts of this knowledge base

Phase 1: Trialled the use of forgetting-based extraction methods for SNOMED CT
(Chen, Alghamdi, S, Walther, Gao, KCAP19)

Phase 2: Developed bespoke knowledge base extraction software for SNOMED CT
(Alghamdi, S, Del-Pinto, Gao, KCAP21; Del-Pinto, S, Gao, ESWC 2022)



The medical knowledge base SNOMED CT

Is one the biggest success stories of symbolic AI

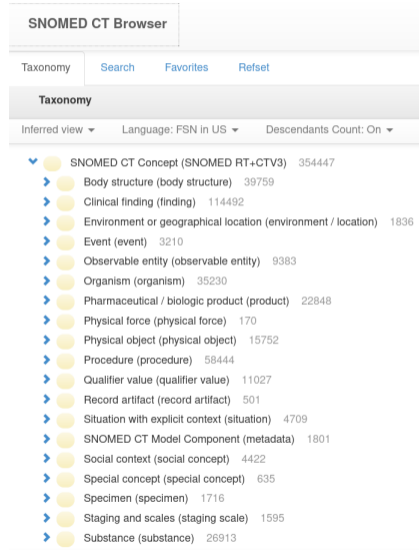
Is a knowledge base of definitions of clinical codes/concepts and their relationships to each other

Provides a global standard in medicine, used in > 50 countries; must be used by doctors/nurses to ensure consistent data entry; used in administrative, reporting, planning, auditing, data analytics, procedural tools

Covers 21 domains defined by > 350,000 precise symbolic statements

Problem:

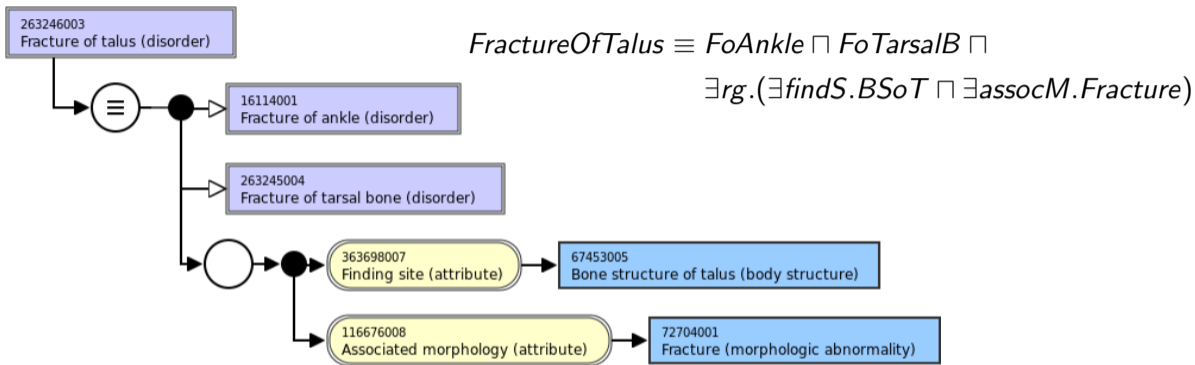
Not all this information in SNOMED CT is needed in every use case



The screenshot shows the SNOMED CT Browser interface. At the top, there is a search bar and navigation tabs for 'Taxonomy', 'Search', 'Favorites', and 'Refset'. Below the tabs, there is a 'Taxonomy' section with a dropdown menu for 'Inferred view', a language selector for 'Language: FSN in US', and a 'Descendants Count: On' dropdown. The main content area displays a list of 21 SNOMED CT concepts, each with a yellow circular icon, a blue arrow, the concept name, and its count. The concepts are: SNOMED CT Concept (SNOMED RT+CTV3) 354447, Body structure (body structure) 39759, Clinical finding (finding) 114492, Environment or geographical location (environment / location) 1836, Event (event) 3210, Observable entity (observable entity) 9383, Organism (organism) 35230, Pharmaceutical / biologic product (product) 22848, Physical force (physical force) 170, Physical object (physical object) 15752, Procedure (procedure) 58444, Qualifier value (qualifier value) 11027, Record artifact (record artifact) 501, Situation with explicit context (situation) 4709, SNOMED CT Model Component (metadata) 1801, Social context (social concept) 4422, Special concept (special concept) 635, Specimen (specimen) 1716, Staging and scales (staging scale) 1595, and Substance (substance) 26913.

SNOMED CT Concept (SNOMED RT+CTV3)	Count
SNOMED CT Concept (SNOMED RT+CTV3)	354447
Body structure (body structure)	39759
Clinical finding (finding)	114492
Environment or geographical location (environment / location)	1836
Event (event)	3210
Observable entity (observable entity)	9383
Organism (organism)	35230
Pharmaceutical / biologic product (product)	22848
Physical force (physical force)	170
Physical object (physical object)	15752
Procedure (procedure)	58444
Qualifier value (qualifier value)	11027
Record artifact (record artifact)	501
Situation with explicit context (situation)	4709
SNOMED CT Model Component (metadata)	1801
Social context (social concept)	4422
Special concept (special concept)	635
Specimen (specimen)	1716
Staging and scales (staging scale)	1595
Substance (substance)	26913

What do definitions of medical concepts in SNOMED CT look like?



The representational language has been carefully designed so that definitions are intuitive

SNOMED CT Editorial Guidelines imply definitions can be neatly visualised

Automated reasoning is tractable, i.e., can be performed in polynomial time

State-of-the-art ELK reasoner developed in the group of Birte Glimm

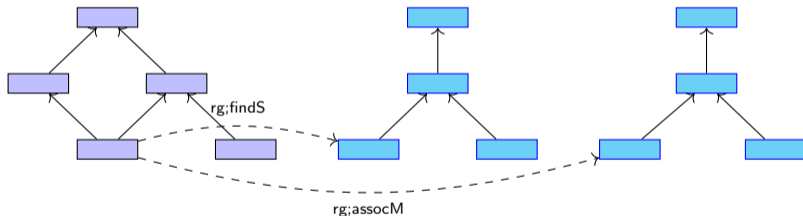
Inferred concept subhierarchies

Using automated reasoning and ELK

It can be inferred which concepts are subtypes of what other concepts

The inferred \sqsubseteq partial order on concepts can be computed

It consists of separate concept subhierarchies linked via attribute relationships

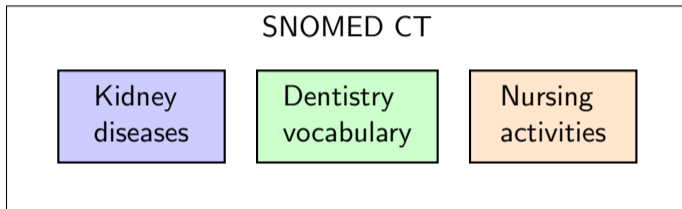


Data from Electronic Health Records can be aggregated and searched at different levels of granularity, e.g., searching for records with FractureOfBone will also return all cases recorded as subtypes, including FractureOfAnkle, FractureOfHand, FractureOfTooth, ...

Their solution to handling the large size of SNOMED CT

Use of reference sets = restricted subsets of concept names

Developed and curated for different tasks and specialist domains by clinical domain experts



Reference sets provide “pick-lists” for easier data entry by clinicians and nurses, to constrain searches/queries, for linkage to other resources, reporting & analytics, ...

But: Reference sets are flat lists of concept names

Better: Have the capability to create stand-alone sub-knowledge bases for such reference sets and support automatic content extraction and sharing for SNOMED CT developers and users

Forgetting-based knowledge base extraction

source $\mathcal{KB} \xrightarrow{-\mathcal{F}} \mathcal{V}$ view

Task: Compute restricted view \mathcal{V} of \mathcal{KB}
by eliminating predicate symbols in \mathcal{F} , the forgetting signature
The user specifies either \mathcal{F} or the keep signature $\mathcal{K} = \text{sig}(\mathcal{KB}) \setminus \mathcal{F}$



Forgetting guarantees:

\mathcal{V} is a new knowledge base that resembles \mathcal{KB} restricted to the keep signature \mathcal{K}
i.e., \mathcal{V} is a sub-knowledge base that can be used instead of \mathcal{KB} for \mathcal{K}

Forgetting has many applications



source \mathcal{KB} $\xrightarrow{-\mathcal{F}}$ \mathcal{V} view

knowledge base/logical theory \longrightarrow restricted view of knowledge base

knowledge base/statement \longrightarrow knowledge base/statement without certain relations

second-order statement \longrightarrow first-order statement

modal axiom \longrightarrow first-order frame condition

logical theory/statement \longrightarrow uniform interpolant

Forgetting offers solutions to many important applications

knowledge base extraction, knowledge base creation,
knowledge base reuse, debugging, repair, information hiding,
semantic difference computation, abductive learning,
agent communication, planning, ...

Automated reasoning can be used to compute forgetting views

... by consequence-finding, i.e., computing all inferred knowledge of a \mathcal{KB} , up to an effective form of redundancy

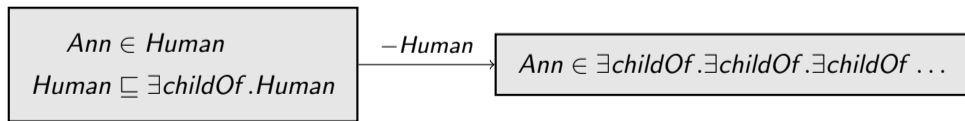
A simple method for computing a forgetting view is to compute all formulas that can be inferred from the given KB and omit those that contain forgetting symbols

$$\mathcal{V} = \{\alpha \mid \alpha \in \mathcal{KB}^*, \text{sig}(\alpha) \cap \mathcal{F} = \emptyset\} \quad \text{where} \quad \mathcal{KB}^* = \{\alpha \mid \mathcal{KB} \models \alpha\} \setminus \text{Red}$$

For propositional logic this set \mathcal{V} is finite \implies forgetting is solvable

However:

Finite forgetting views need not exist for description logic \mathcal{ALC} or the Horn fragment \mathcal{EL}

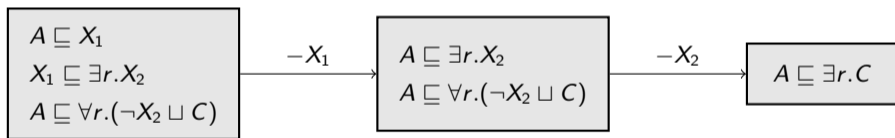


Automated forgetting for ontologies

Key research questions: (i) Devising consequence-finding algorithms for ontologies?
(ii) Is it possible to represent infinite forgetting views compactly?

Our LETHE system

= state-of-the art forgetting-tool for various DLs (Koopmann, S, 2013–2021)
Is guaranteed to terminate and always produces finite views/representations (by use of helper symbols or fixpoint operator); 2EXP size bound which is better than 3EXP theoretical bound



Our FAME system

= forgetting tool based on Ackermann's Lemma (Zhao et al, S, 2015–21)
Always terminates and produces correct views, but not guaranteed to be successful

Results of empirical evaluations have shown excellent timings and success rates for both

Our other forgetting or knowledge base extraction type tools

Abduction hypothesis generator

Based on LETHE

UI-FAME

Hybrid forgetting tool for DLs

LDiff-FAME

DL ontology version tracker

SNOMED CT Subontology builder

Industry collaboration

F3

Fine-grained forgetting prototype

Witness generator

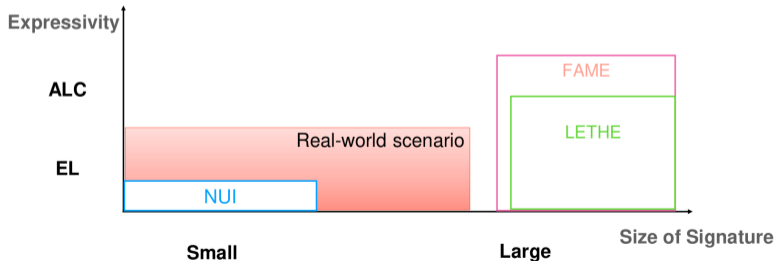
Returns witnesses for forgetting symbols

Trialling forgetting on SNOMED CT (Phase 1)

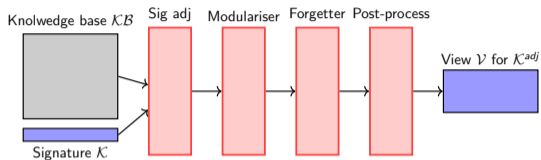
Previous evaluations of our LETHE and FAME forgetting tools had shown excellent results (success rates and performance) on medium size knowledge bases and large keep signatures

The tools are not suited for the SNOMED CT use case, where the aim is to compute knowledge base extracts for very small real-world signatures ($< 1\%$)

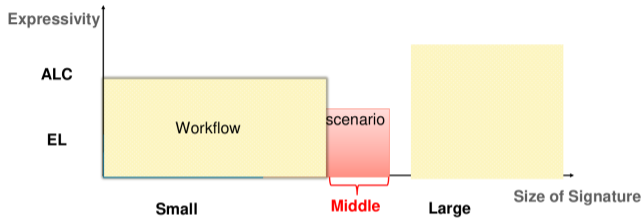
NUI system can handle very small keep signatures but struggles with larger numbers



Signature adjustment \longrightarrow modularisation \longrightarrow forgetting workflow



Modularisation is itself a knowledge base extraction method, but modules are often too large



Finding:

Workflow allows feasible sub-knowledge base extraction on small and large keep signatures
(Chen, Alghamdi, S, Walther, Gao, KCAP19)

But: More work is needed to create extracts suitable for SNOMED CT use cases and tooling

Bespoke subontology builder software for SNOMED CT (Phase 2)

Input: A set of **focus concept names**, e.g., kidney diseases reference set

Output: A **concise, reusable subontology** \mathcal{S} containing content relevant to focus concepts, that can be used instead of SNOMED CT and satisfies SNOMED CT Editorial Guidance

Method:

- ▶ Uses **definition-oriented consequence-based knowledge extraction**, tailored to the needs of SNOMED CT use cases and tooling
- ▶ Different to forgetting views, subontologies can contain concept names other than focus concept names, called **supporting concept names**
Definitions of the focus concept names in \mathcal{S} are equivalent to definitions in SNOMED CT
 \mathcal{S} may include just enough supporting concept names for these definitions
The concept hierarchy of \mathcal{S} is the same as that of SNOMED CT for the signature of \mathcal{S}
- ▶ Semantic completeness only for definitions of focus concept names and concept name hierarchy of \mathcal{S}

Evaluation

Focus set	Focus set size	Subontology size		Module size		Time
	Concepts	Statements	Concepts	Statements	Concepts	Secs
ERA-EDTA	184	797	786	3076	3086	51
Dentistry	226	648	642	1449	1478	50
Nursing	1337	2638	2616	5579	5708	59
Orphanet	5681	10666	10637	27595	27625	113
IPS	8182	13987	13915	53736	53708	124
MP	21732	28342	27293	29386	28867	137
GPS	26159	35828	35733	86167	86374	266

Excellent performance, a lot less reasoning is required

In subontologies the information is greatly compressed

For IPS: SNOMED CT 350,000 → Module 54,000 → Subontology 14,000

A lot fewer concept definitions, less distraction and the subontology is easier to navigate

Clinical concept definition of 'Acute renal failure' in SNOMED CT

Parents

- Acute renal impairment (disorder)
- Renal failure syndrome (disorder)

Acute renal failure syndrome (disorder) ☆

SCTID: 14669001

14669001 | Acute renal failure syndrome (disorder) |

- en Acute renal failure syndrome (disorder)
- en Acute renal failure syndrome
- en AKI - acute kidney injury
- en ARF - Acute renal failure
- en Acute renal failure

Interprets → Measurement of renal function
Has interpretation → Impaired

Finding site → Kidney structure

Clinical course → Sudden onset AND/OR short duration

Children (27)

- Acute kidney failure stage 1 (disorder)
- Acute kidney failure stage 2 (disorder)
- Acute kidney failure stage 3 (disorder)
- Acute kidney injury due to circulatory failure (disorder)
- Acute kidney injury due to disease caused by severe acute respiratory syndrome coronavirus 2 (disorder)
- Acute kidney injury due to nephrotoxicity (disorder)
- Acute kidney injury due to sepsis (disorder)
- Acute kidney injury following administration of contrast media (disorder)
- Acute milk alkali syndrome (disorder)
- Acute renal failure due to acute cortical necrosis (disorder)
- Acute renal failure due to ischemia (disorder)
- Acute renal failure due to obstruction (disorder)
- Acute renal failure due to procedure (disorder)
- Acute renal failure following molar AND/OR ectopic pregnancy (disorder)
- Acute renal failure on dialysis (disorder)
- Acute renal failure with oliguria (disorder)
- Acute renal papillary necrosis with renal failure (disorder)
- Acute-on-chronic renal failure (disorder)
- Crush syndrome (disorder)
- Failed attempted abortion with acute renal failure (disorder)
- Hepatorenal syndrome (disorder)

Focus concept definition of the same concept in ERA-EDTA Subontology

Parents

- Renal failure syndrome (disorder)

Acute renal failure syndrome (disorder) ☆

SCTID: 14669001

14669001 | Acute renal failure syndrome (disorder) |

- en Acute renal failure syndrome (disorder)
- en Acute renal failure syndrome
- en ARF - Acute renal failure
- en Acute renal failure
- en Acute renal failure syndrome (disorder)

Has interpretation → Impaired
Interprets → Measurement of renal function

Clinical course → Sudden onset AND/OR short duration

Finding site → Kidney structure

Children (6)

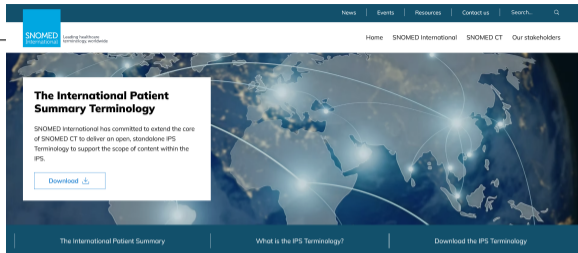
- Acute kidney injury due to circulatory failure (disorder)
- Acute kidney injury due to hypovolemia (disorder)
- Acute kidney injury due to sepsis (disorder)
- Crush syndrome (disorder)
- Hepatorenal syndrome (disorder)
- Nephrotoxic acute renal failure (disorder)

Typical comment in user evaluation:
“Restriction to a speciality subontology makes navigation and content finding easier”

Subontologies are available to the global SNOMED community as a new functionality and are already used in a couple of projects

International Patient Summary subontology was created and released as a new product needed for across boundaries healthcare information, addressing an international standardization effort (ISO, CEN, GDHP); no licensing restrictions
(Del-Pinto, S, Gao, Alghamdi, Lopez Osornio, Roy, MedInfo 2023)

Our subontology builder tool is used at SNOMED Intl in several **content development and quality assurance projects** (substance, anatomy)



Summary

Focused on two main knowledge base extraction methods

Forgetting:

- ▶ Powerful, versatile method with many applications in different domains
- ▶ Automation is possible, despite the problems being generally hard

Subontology generation:

- ▶ Tailored to the needs of the medical knowledge base SNOMED CT
- ▶ Subontologies contain just the right amount of information to satisfy the requirements of SNOMED Intl and they scale

Ongoing:

- ▶ Abductive learning
- ▶ Witness computation for \mathcal{F} -symbols
- ▶ Extraction of deductive forgetting views from semantic forgetting views
- ▶ Systematic comparison of KB extraction methods
- ▶ Combining subontologies and semantic difference computation to compare and merge the modelling in the same subdomain of two medical knowledge bases

There are lots of opportunities for further work and interesting PhD projects!

Thanks!

Patrick Koopmann, Yizheng Zhao, Mohammad Khodadadi, Giorgos Stoilos, Damir Juric, Ghadah Alghamdi, Jieying Chen, Dirk Walther, Warren Del-Pinto, Mostafa Sakr, Ruba Alassaf, Yongsheng Gao, Kai Kewley, Rory Davidson, Jim Case, Monica Harry, Suzy Roy, Alejandro Lopez Osornio and many others

Some tools/resources:

<https://github.com/IHTSDO/snomed-subontology-extraction>

<https://iaa.snomed.tools>

<https://www.snomed.org/international-patient-summary-terminology>

Contact: Renate.Schmidt@manchester.ac.uk

Funded by:



Thank you !
Questions ?