Biomedical Informatics

Lefteris Koumakis

Medical Ontologies

Medical Ontologies – Why?

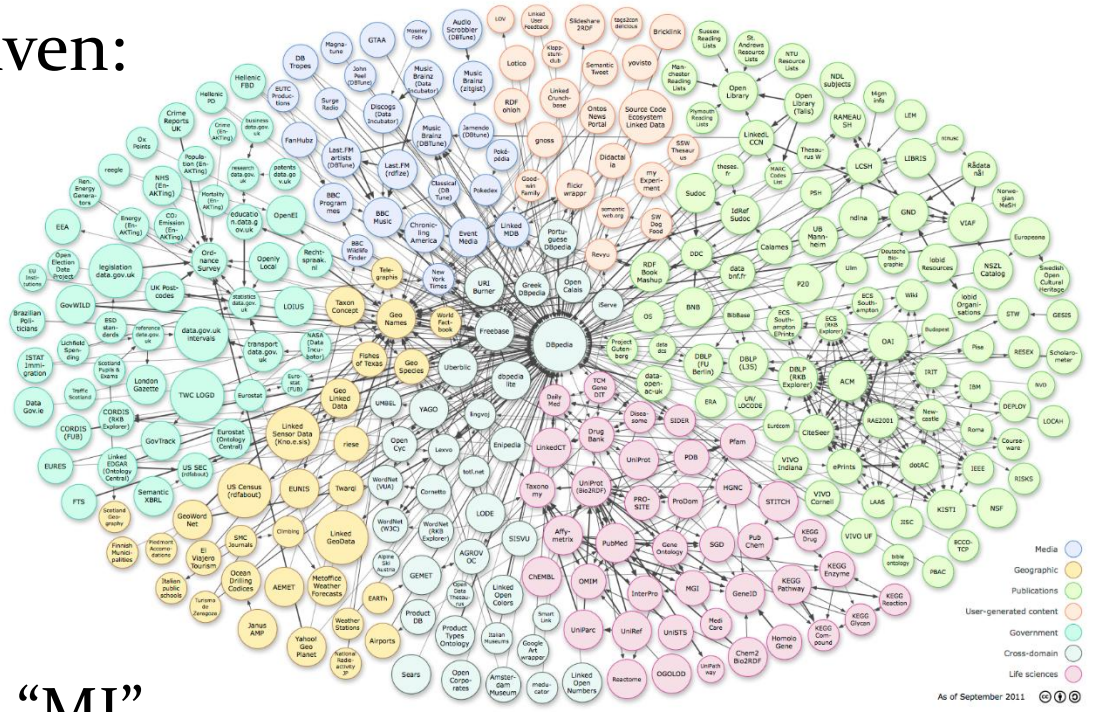
Ontologies are a way to represent our knowledge on a specific topic

- Modern healthcare is increasingly data-driven:
  - Electronic Health Records (EHRs)
  - AI/ML diagnostic systems
  - Biomedical signal processing
  - Interoperable health platforms

Medical data is messy, heterogeneous

- Example:
  - “Heart attack” vs “Myocardial infarction” vs “MI”

Without standardization, systems cannot communicate effectively.



Terminology vs Ontology

Terminology

A **controlled vocabulary** of medical terms.

- Focus: naming and coding
- Example: assigning a code to “diabetes”

Ontology

Structured representation of knowledge

- Defines:
  - Concepts (entities)
  - Relationships
  - Constraints

Machine-readable

A **graph-based model of knowledge**

Feature	Terminology	Ontology
Structure	Flat / hierarchical	Graph with rich relations
Purpose	Standard naming	Knowledge representation
Example	ICD codes	SNOMED CT

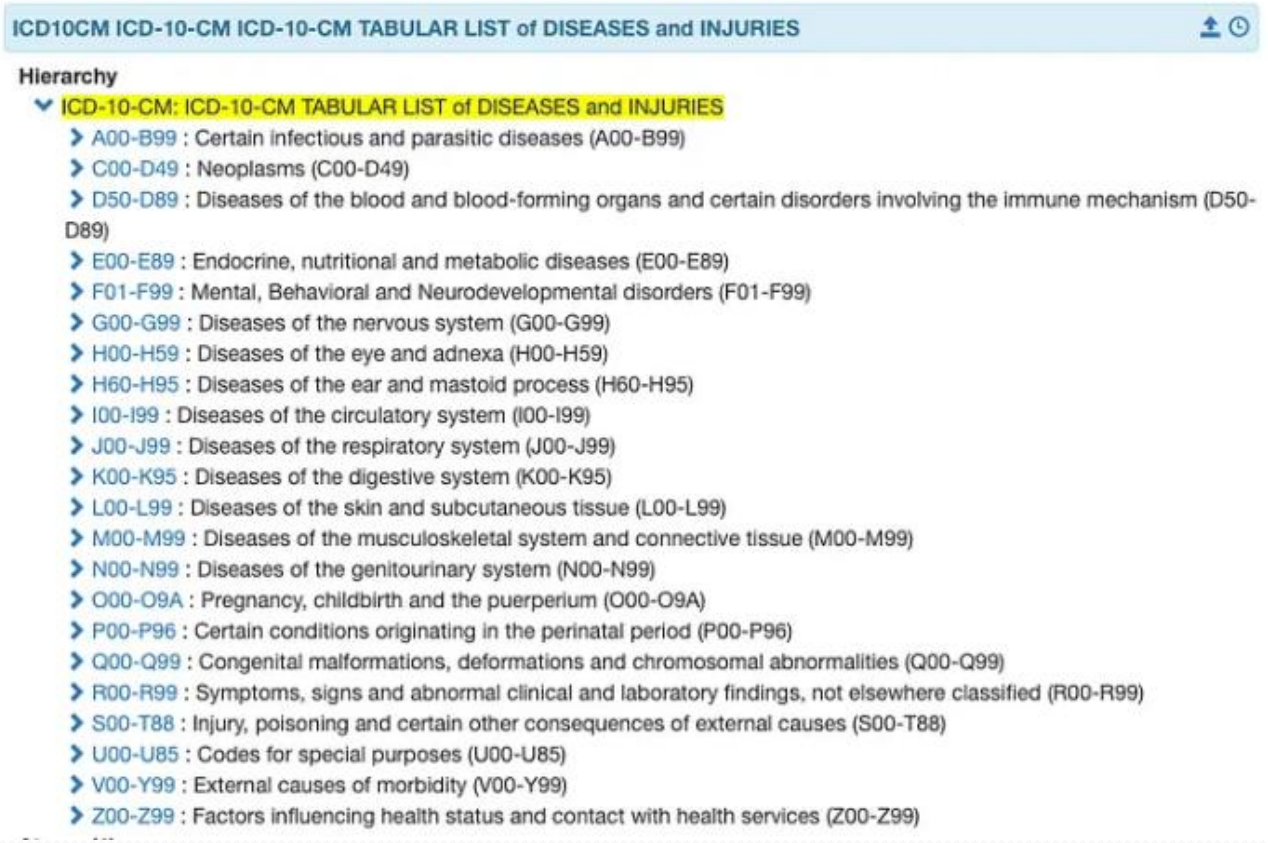
Terminology example - ICD

ICD (International Classification of Diseases)

- Used for:
  - Epidemiology

- Example:
  - E11 → Type 2 Diabetes

Simple hierarchy, not deeply expressive



Ontology example - SNOMED

SNOMED CT (Systematized Nomenclature of Medicine)

- One of the most comprehensive systems
- Contains:
  - Diseases
  - Procedures
  - Findings
- Supports relationships like:
  - “is-a”
  - “part-of”
  - “associated-with”

Ontologies

Ontology is a core and critical area of philosophy

- Specifically, metaphysics (describing what exists and categories of existence)

In computer science, ontologies originated in the artificial intelligence community

- Computers needed to understand human logic and decision-making

Ontologies are a way to represent our knowledge on a specific topic

Ontologies allow us to share information using a common language

Ontologies help computers “understand” a subject and apply logic

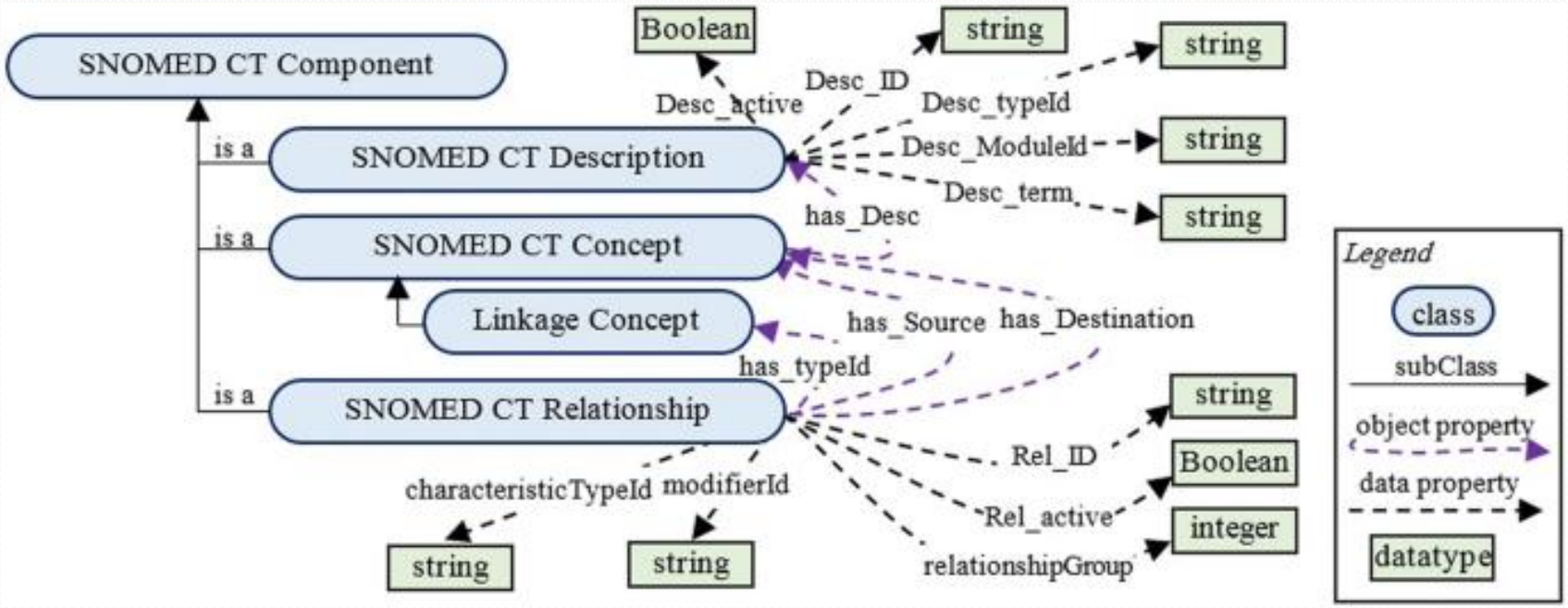
Ontology Graph

How Ontologies Work

- Nodes = concepts
- Edges = relationships
- Stored as graphs

Supports:

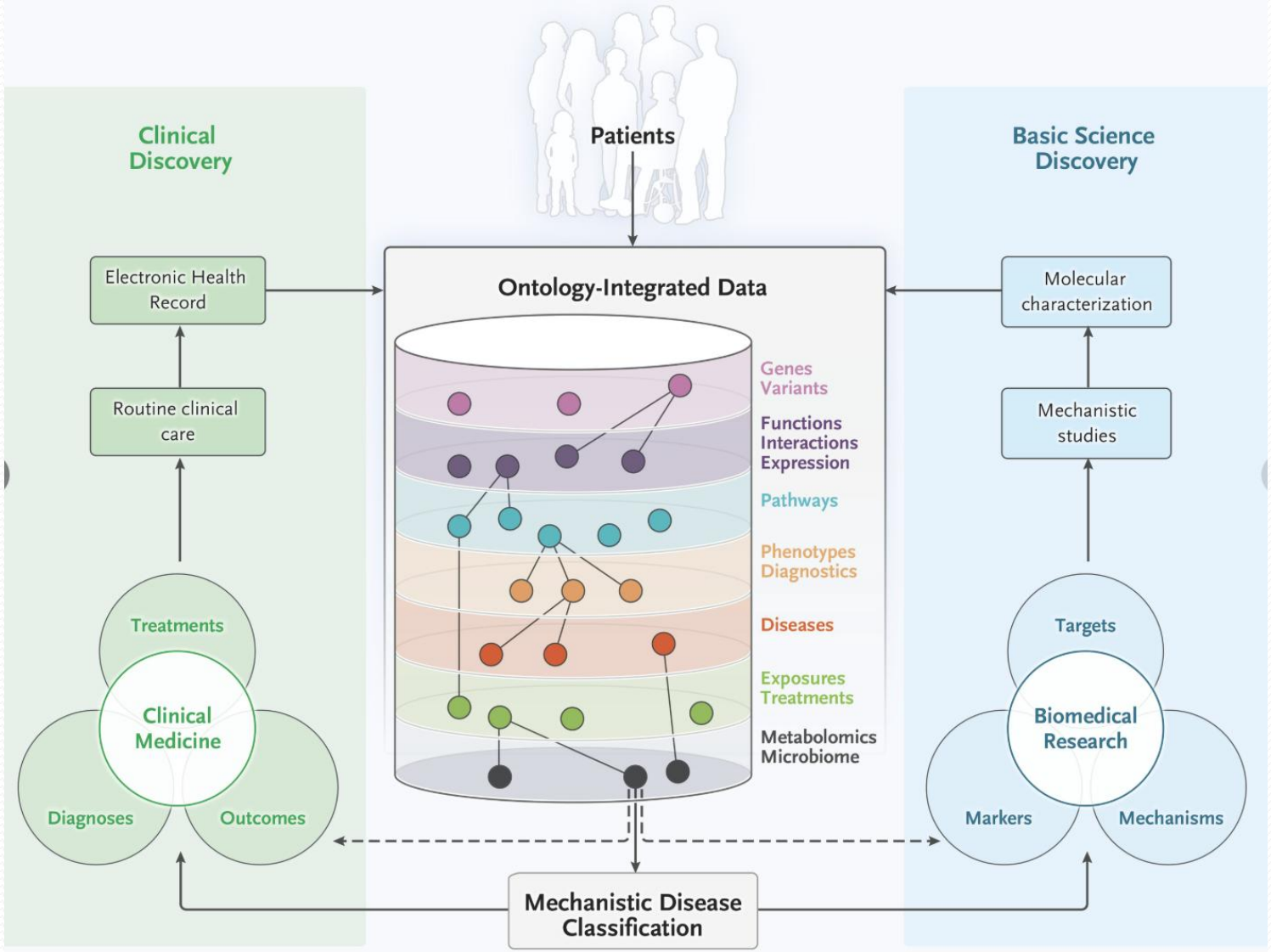
- Inference
- Reasoning
- Knowledge linking



The power of ontologies

- **Graph Structure**
- Medical ontologies are typically modeled as **graphs**:
- Nodes → concepts
- Edges → relationships
- Example:
- Pneumonia → *is-a* → Lung disease
- Lung → *part-of* → Respiratory system

The big picture

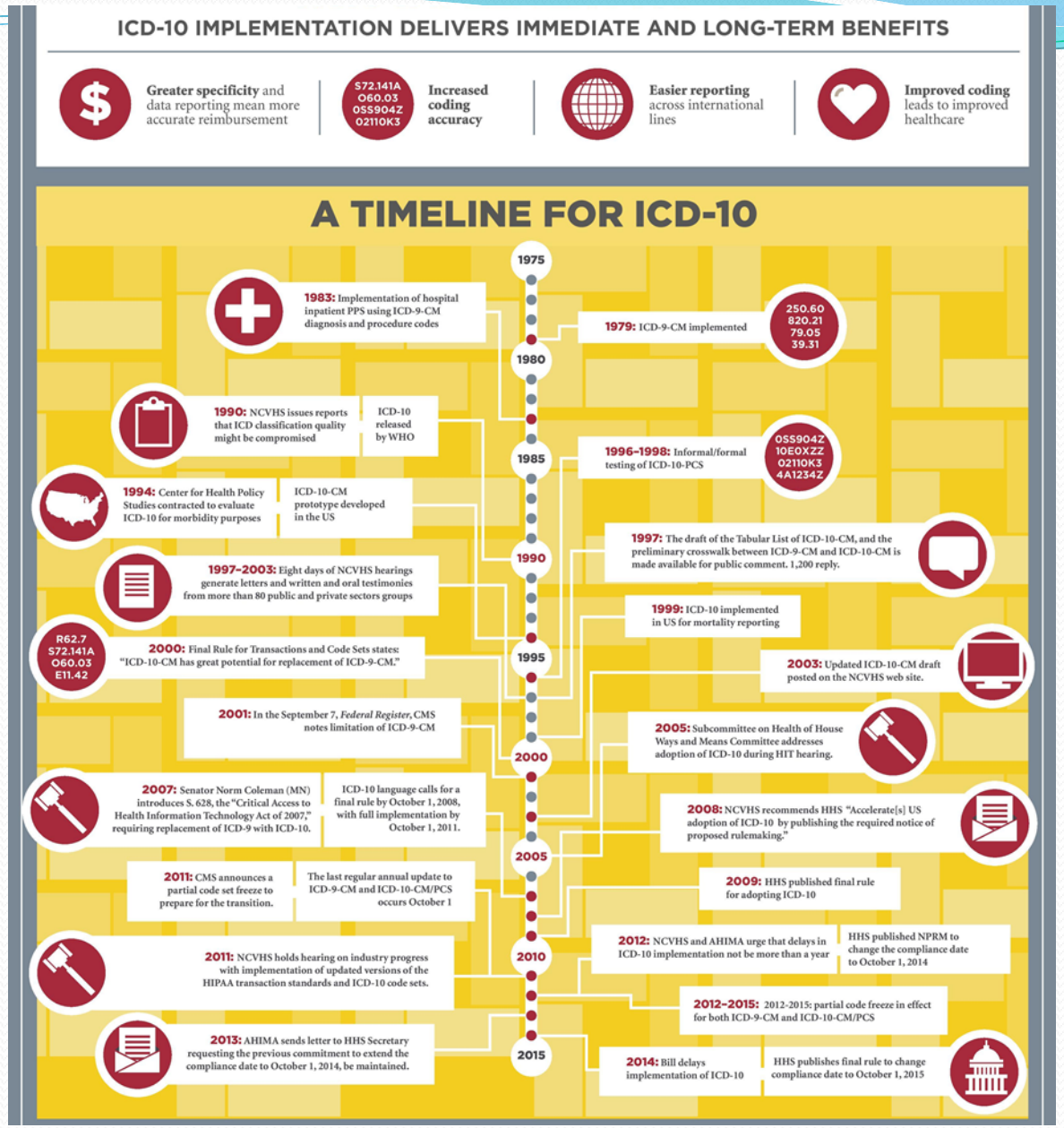


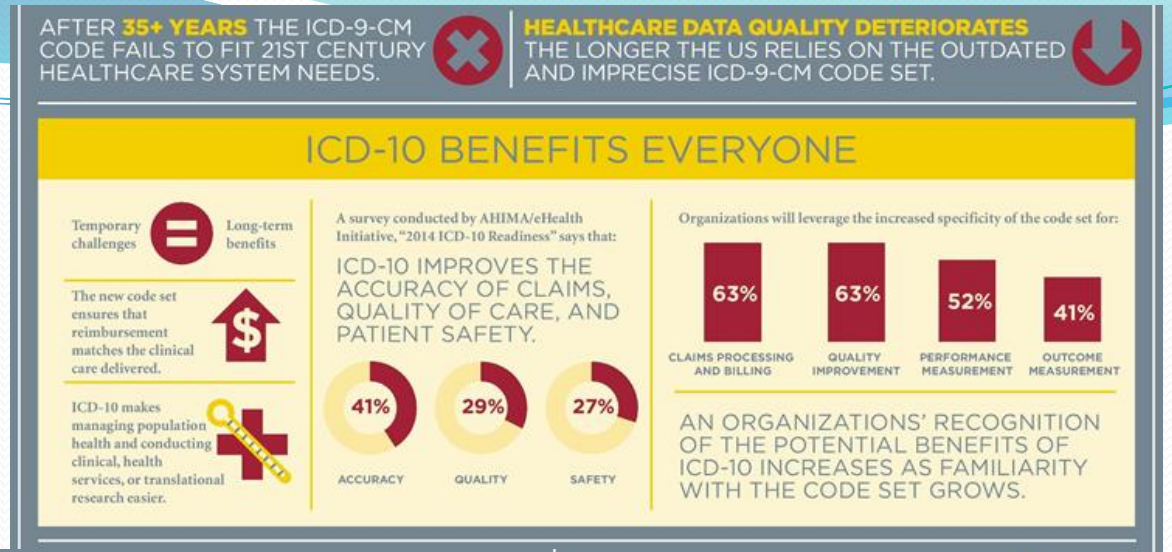
Biomedical ontologies

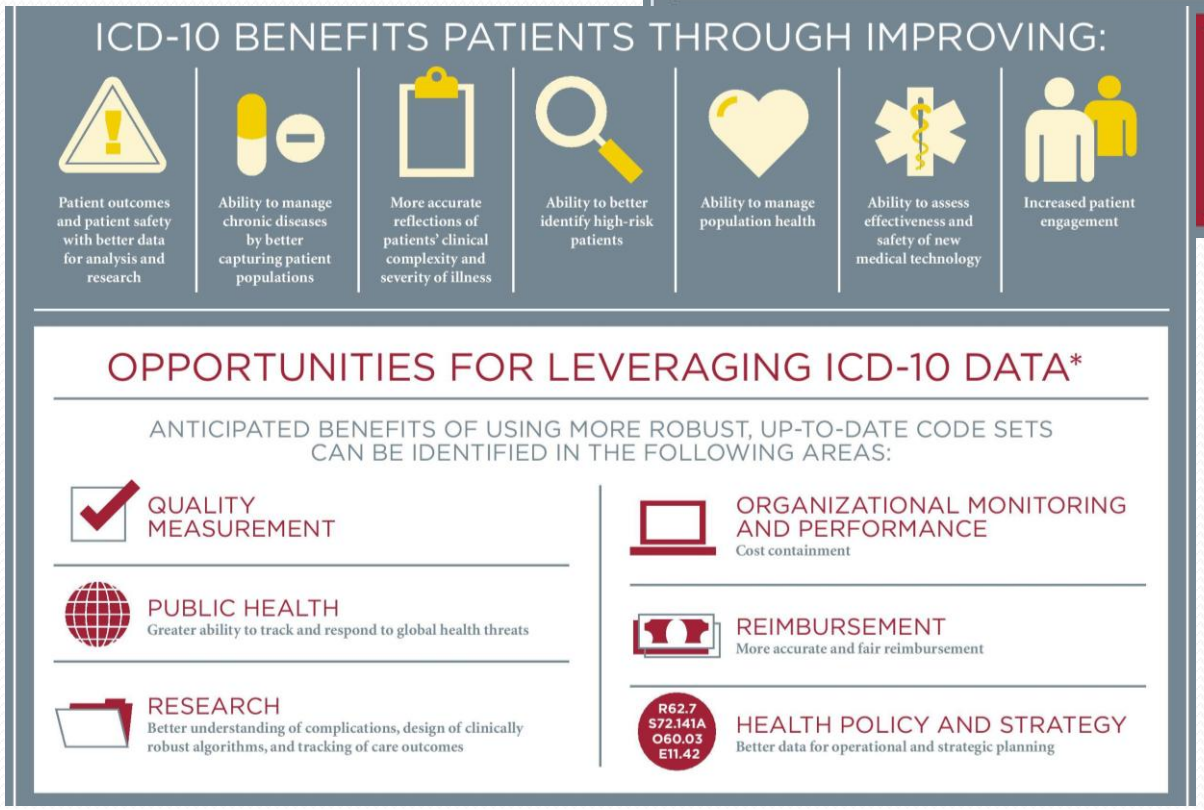
International Classification of Diseases (ICD)

- International Classification of Diseases
- The first international classification edition, known as the International List of Causes of Death, was adopted by the International Statistical Institute in 1893.

- WHO was entrusted with the ICD at its creation in 1948 and published the 6th version, ICD-6, that incorporated morbidity for the first time.


ICD-10 IMPLEMENTATION DELIVERS IMMEDIATE AND LONG-TERM BENEFITS
Greater specificity and data reporting mean more accurate reimbursement
S72.141A O60.03 0SS904Z 02110K3
Increased coding accuracy
Easier reporting across international lines
Improved coding leads to improved healthcare
A TIMELINE FOR ICD-10
1975
1983: Implementation of hospital inpatient PPS using ICD-9-CM diagnosis and procedure codes
1979: ICD-9-CM implemented
250.60 820.21 79.05 39.31
1980
1990: NCVHS issues reports that ICD classification quality might be compromised
ICD-10 released by WHO
1985
1996–1998: Informal/formal testing of ICD-10-PCS
0SS904Z 10E0XZZ 02110K3 4A1234Z
1994: Center for Health Policy Studies contracted to evaluate ICD-10 for morbidity purposes
ICD-10-CM prototype developed in the US
1997: The draft of the Tabular List of ICD-10-CM, and the preliminary crosswalk between ICD-9-CM and ICD-10-CM is made available for public comment. 1,200 reply.
1990
1997–2003: Eight days of NCVHS hearings generate letters and written and oral testimonies from more than 80 public and private sectors groups
1999: ICD-10 implemented in US for mortality reporting
R62.7 S72.141A O60.03 E11.42
2000: Final Rule for Transactions and Code Sets states: "ICD-10-CM has great potential for replacement of ICD-9-CM."
1995
2003: Updated ICD-10-CM draft posted on the NCVHS web site.
2001: In the September 7, Federal Register, CMS notes limitation of ICD-9-CM
2005: Subcommittee on Health of House Ways and Means Committee addresses adoption of ICD-10 during HIT hearing.
2000
2007: Senator Norm Coleman (MN) introduces S. 628, the "Critical Access to Health Information Technology Act of 2007," requiring replacement of ICD-9 with ICD-10.
ICD-10 language calls for a final rule by October 1, 2008, with full implementation by October 1, 2011.
2008: NCVHS recommends HHS "Accelerate[s] US adoption of ICD-10 by publishing the required notice of proposed rulemaking."
2005
2011: CMS announces a partial code set freeze to prepare for the transition.
The last regular annual update to ICD-9-CM and ICD-10-CM/PCS occurs October 1
2009: HHS published final rule for adopting ICD-10
2011: NCVHS holds hearing on industry progress with implementation of updated versions of the HIPAA transaction standards and ICD-10 code sets.
2012: NCVHS and AHIMA urge that delays in ICD-10 implementation not be more than a year
HHS published NPRM to change the compliance date to October 1, 2014
2010
2012–2015: 2012-2015: partial code freeze in effect for both ICD-9-CM and ICD-10-CM/PCS
2013: AHIMA sends letter to HHS Secretary requesting the previous commitment to extend the compliance date to October 1, 2014, be maintained.
2014: Bill delays implementation of ICD-10
HHS publishes final rule to change compliance date to October 1, 2015
2015

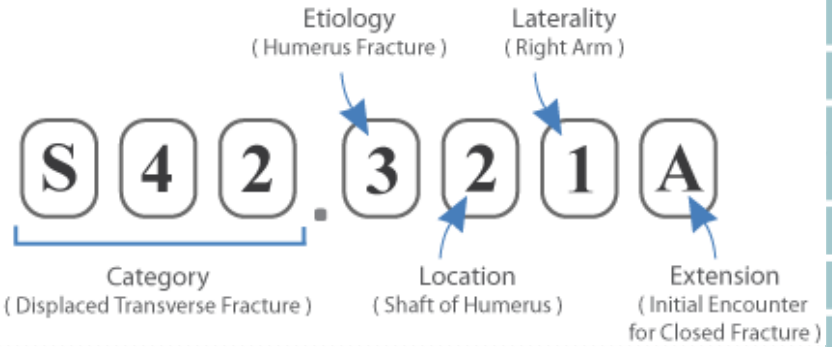

AFTER 35+ YEARS THE ICD-9-CM CODE FAILS TO FIT 21ST CENTURY HEALTHCARE SYSTEM NEEDS.
HEALTHCARE DATA QUALITY DETERIORATES
THE LONGER THE US RELIES ON THE OUTDATED AND IMPRECISE ICD-9-CM CODE SET.
ICD-10 BENEFITS EVERYONE
Temporary challenges
Long-term benefits
The new code set ensures that reimbursement matches the clinical care delivered.
ICD-10 makes managing population health and conducting clinical, health services, or translational research easier.
A survey conducted by AHIMA/eHealth Initiative, "2014 ICD-10 Readiness" says that:
ICD-10 IMPROVES THE ACCURACY OF CLAIMS, QUALITY OF CARE, AND PATIENT SAFETY.
41%
29%
27%
ACCURACY
QUALITY
SAFETY
Organizations will leverage the increased specificity of the code set for:
63%
63%
52%
41%
CLAIMS PROCESSING AND BILLING
QUALITY IMPROVEMENT
PERFORMANCE MEASUREMENT
OUTCOME MEASUREMENT
AN ORGANIZATIONS' RECOGNITION OF THE POTENTIAL BENEFITS OF ICD-10 INCREASES AS FAMILIARITY WITH THE CODE SET GROWS.


ICD-10 BENEFITS PATIENTS THROUGH IMPROVING:
Patient outcomes and patient safety with better data for analysis and research
Ability to manage chronic diseases by better capturing patient populations
More accurate reflections of patients' clinical complexity and severity of illness
Ability to better identify high-risk patients
Ability to manage population health
Ability to assess effectiveness and safety of new medical technology
Increased patient engagement
OPPORTUNITIES FOR LEVERAGING ICD-10 DATA*
ANTICIPATED BENEFITS OF USING MORE ROBUST, UP-TO-DATE CODE SETS CAN BE IDENTIFIED IN THE FOLLOWING AREAS:
QUALITY MEASUREMENT
PUBLIC HEALTH
Greater ability to track and respond to global health threats
RESEARCH
Better understanding of complications, design of clinically robust algorithms, and tracking of care outcomes
ORGANIZATIONAL MONITORING AND PERFORMANCE
Cost containment
REIMBURSEMENT
More accurate and fair reimbursement
R62.7
S72.141A
O60.03
E11.42
HEALTH POLICY AND STRATEGY
Better data for operational and strategic planning


• Research efforts for population health
• Public health reporting
• Successful adaptation to changing reimbursement methods
ICD-10

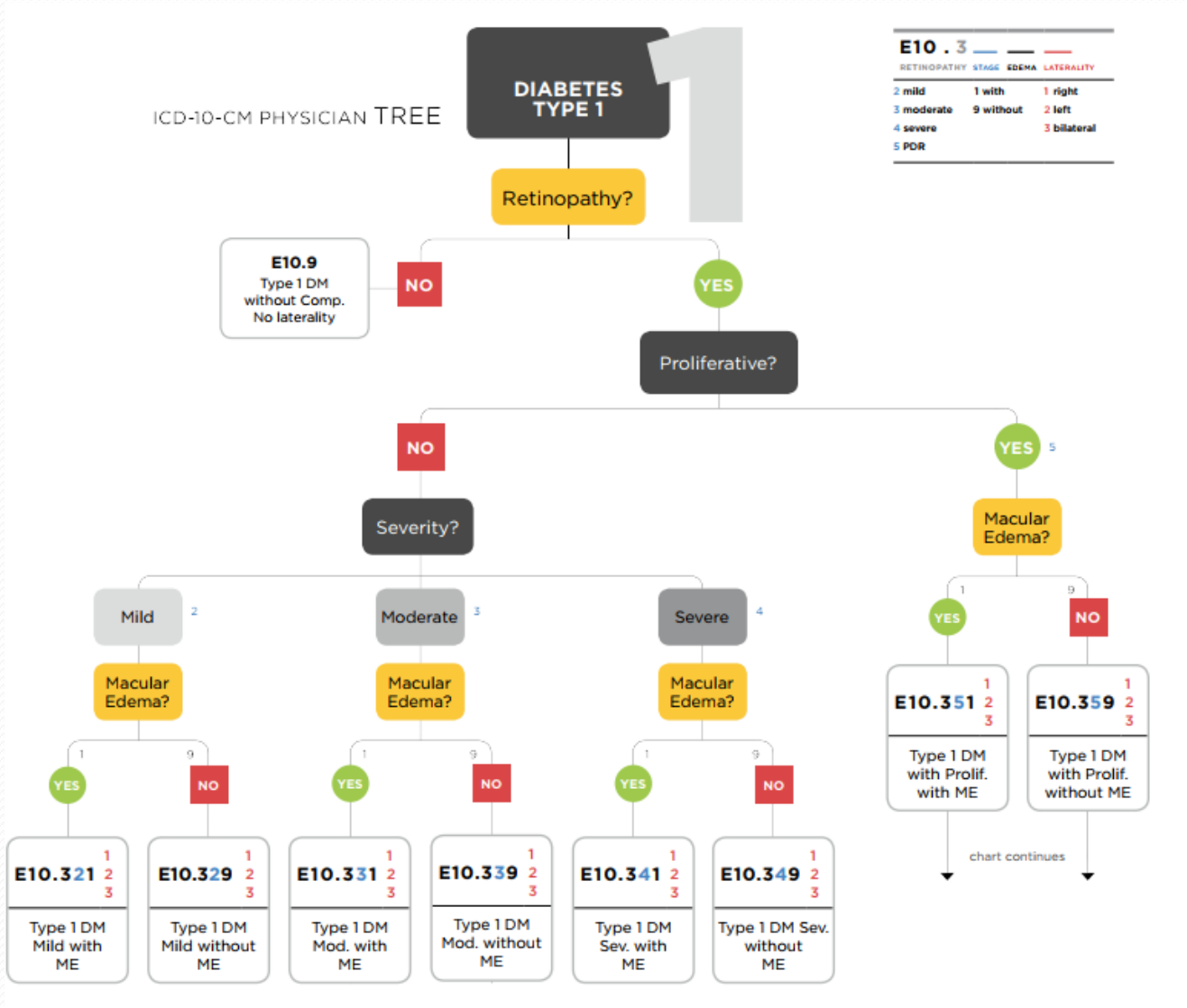
Differences Between ICD-9-CM and ICD-10 Code Sets		
	ICD-9-CM	**ICD-10 code sets**
Procedure	3,824 codes	71,924 codes
Diagnosis	14,025 codes	69, 823 codes

ICD-10 Code Structure Changes (selected details)		
	Old	**New**
Diagnosis Structure	ICD-9-CM • 3 -5 characters • First character is numeric or alpha • Characters 2-5 are numeric	ICD-10-CM • 3 -7 characters • Character 1 is alpha • Character 2 is numeric • Characters 3 – 7 can be alpha or numeric
Procedure Structure	ICD-9-CM • 3-4 characters • All characters are numeric • All codes have at least 3 characters	ICD-10-PCS • ICD-10-PCS has 7 characters • Each can be either alpha or numeric • Numbers 0-9; letters A-H, J-N, P-Z



I	Λοιμώδη και παρασιτικά νοσήματα
II	Νεοπλασίες
III	Ασθένειες του αίματος και των αιμοποιητικών οργάνων, καθώς και διαταραχές που αναφέρονται σε ανοσοποιητικούς μηχανισμούς
IV	Ενδοκρινικά, διατροφικά και μεταβολικά νοσήματα
V	Διαταραχές νόησης και συμπεριφοράς
VI	Ασθένειες του νευρικού συστήματος
VII	Οφθαλμολογικές ασθένειες
VIII	Ασθένειες των ώτων και της μαστοειδούς διαδικασίας
IX	Ασθένειες του κυκλοφορικού συστήματος
X	Ασθένειες του αναπνευστικού συστήματος
XI	Ασθένειες του πεπτικού συστήματος
XII	Ασθένειες του δέρματος και του υποδόριου ιστού
XIII	Ασθένειες του μυοσκελετικού συστήματος και του συνδετικού ιστού
XIV	Ασθένειες του ουροποιητικού συστήματος
XV	Εγκυμοσύνη, τοκετός και επιλόχειες επιπλοκές
XVI	Ανωμαλίες προερχόμενες από την προγενετική περίοδο
XVII	Συγγενείς παραμορφώσεις και χρωμοσωμικές ανωμαλίες
XVIII	Συμπτώματα, σημεία και παθολογικά κλινικά ή εργαστηριακά ευρήματα μη ταξινομημένα σε ειδικές κατηγορίες
XIX	Τραύματα, δηλητηριάσεις και συνέπειες άλλων εξωτερικών παραγόντων

Subspecialty ICD-10 Decision Trees and Guides

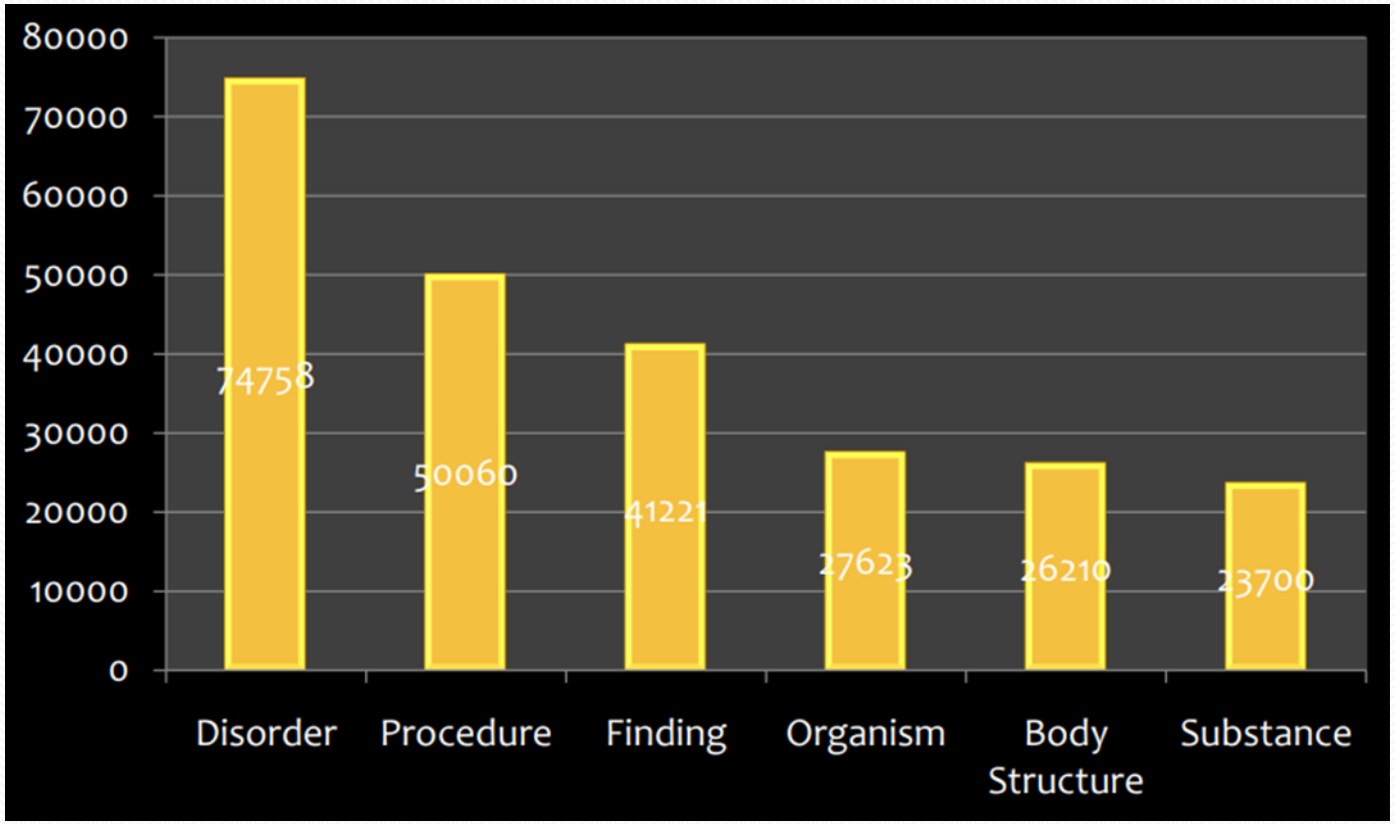


Snomed CT

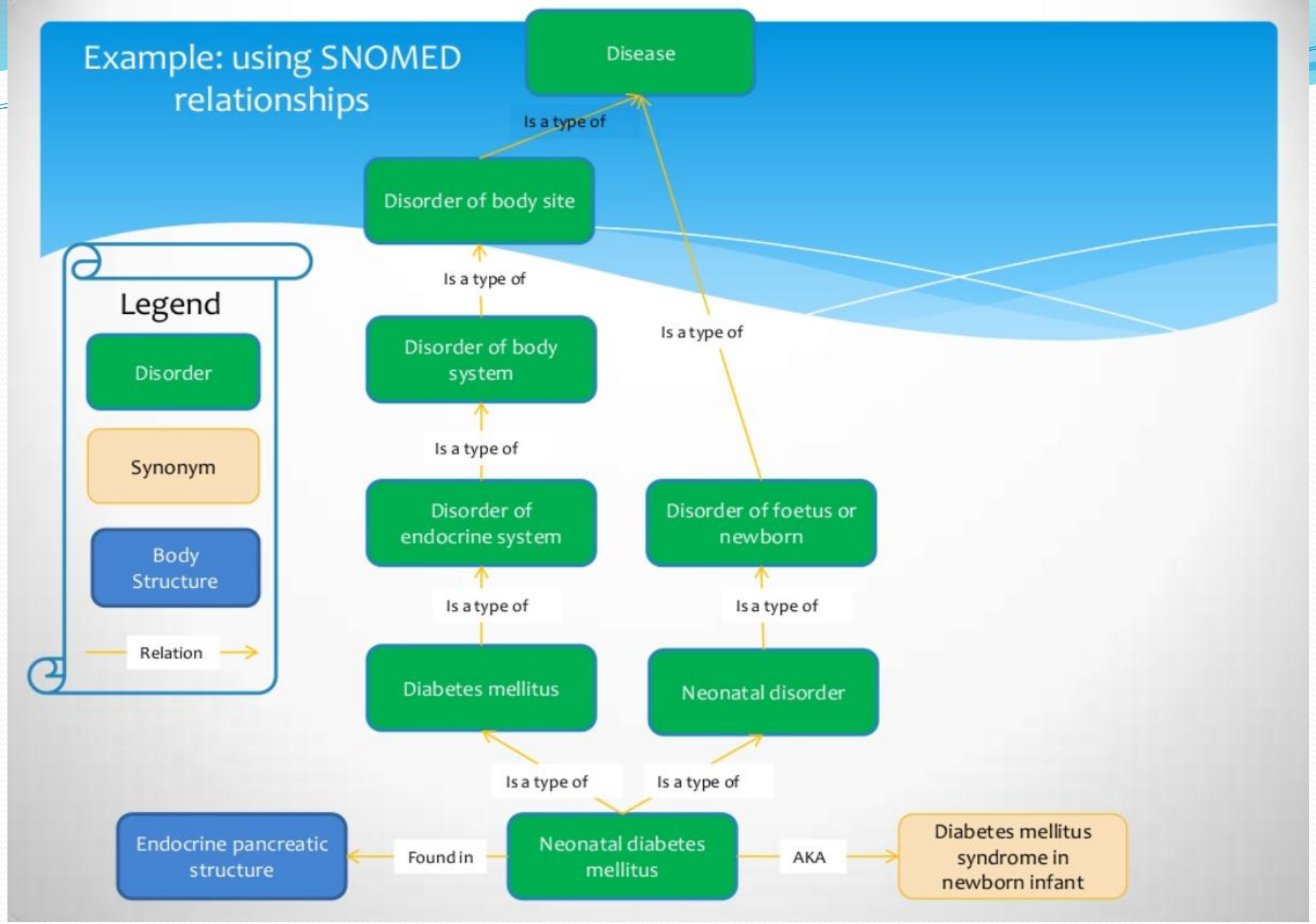
SYSTEMATIZED NOMENCLATURE OF MEDICINE - CLINICAL TERMS

- healthcare terminology including comprehensive coverage of diseases, clinical findings, therapies, procedures and outcomes.
- It provides the core general terminology for the electronic health record
- contains more than **357,000** concepts with unique meanings and formal logic-based definitions organised into hierarchies.
- Is the most comprehensive, multilingual clinical healthcare terminology in the world
- Is a resource with comprehensive, scientifically validated clinical content
- Enables consistent, processable representation of clinical content in electronic health records
- Is mapped to other international standards
- Is already used in more than fifty countries

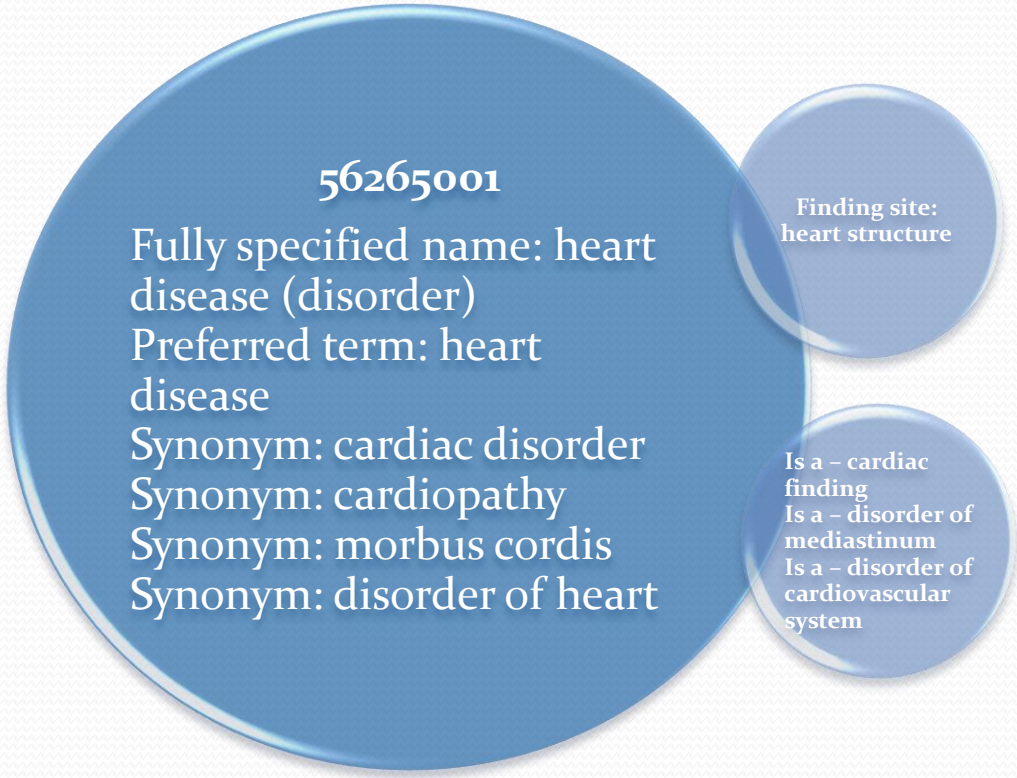
Top Snomed Hierarchies

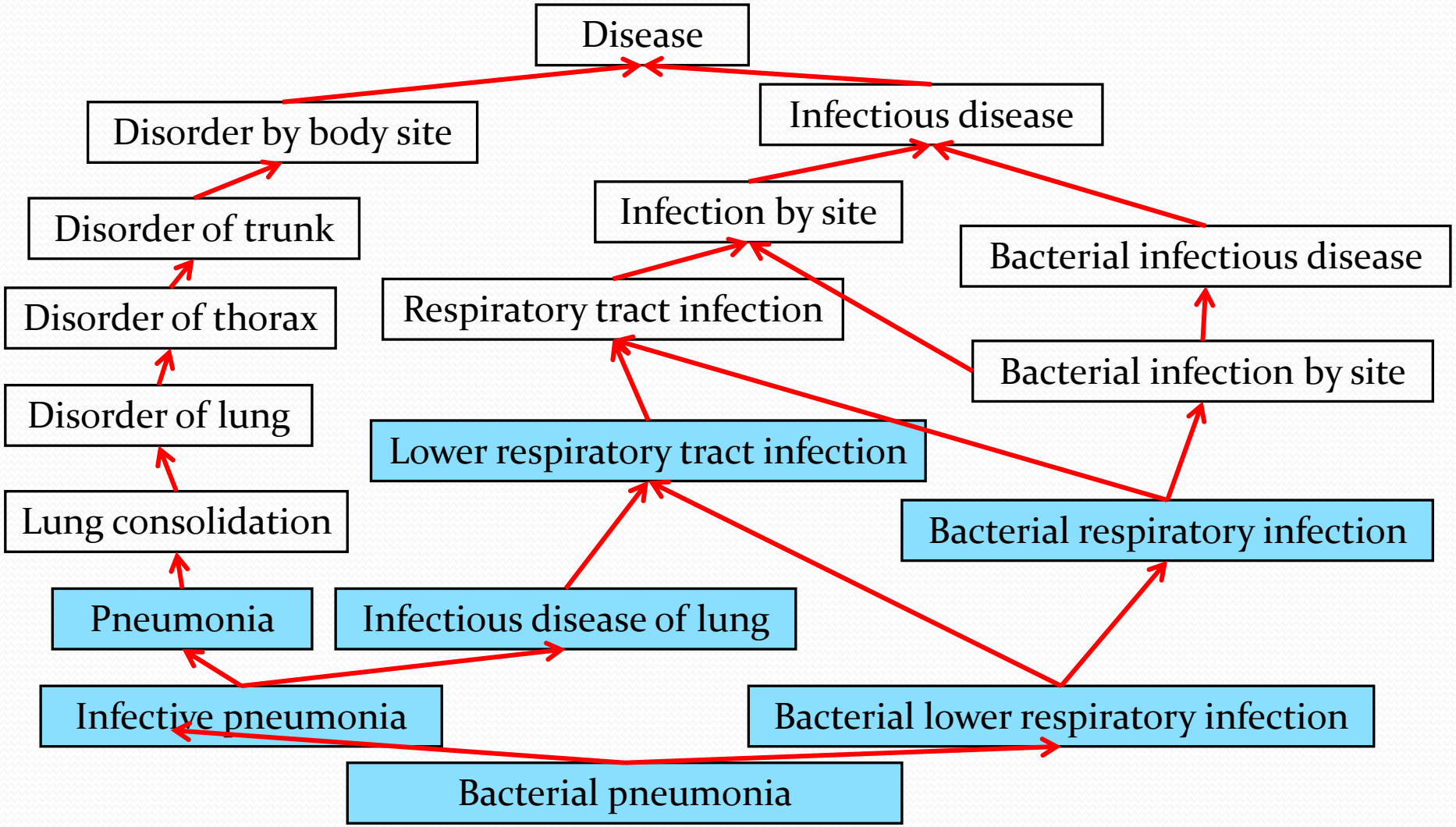


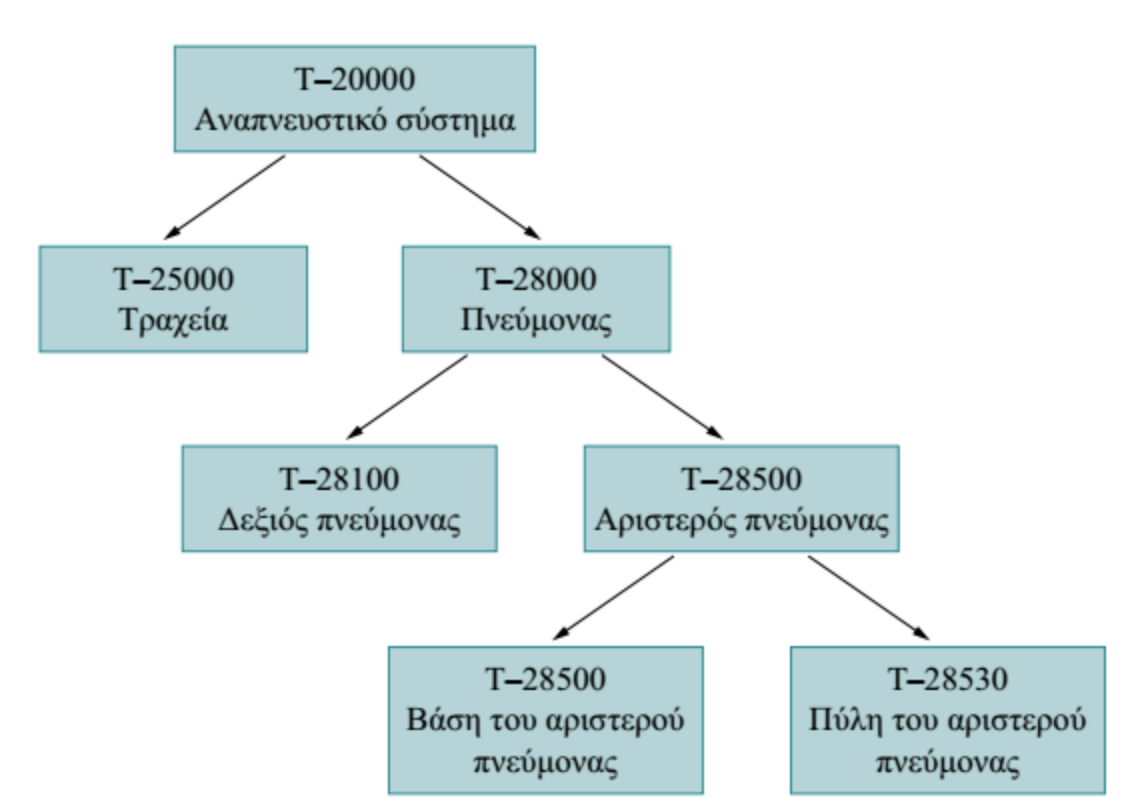
http://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=en-edition&release=v20170731&server=http://browser.ihtsdotools.org/api/v1/snomed&langRefset=900000000000509007


Example: using SNOMED relationships
Disease
Is a type of
Disorder of body site
Is a type of
Disorder of body system
Is a type of
Disorder of endocrine system
Is a type of
Diabetes mellitus
Is a type of
Endocrine pancreatic structure
Found in
Neonatal diabetes mellitus
AKA
Diabetes mellitus syndrome in newborn infant
Is a type of
Disorder of foetus or newborn
Is a type of
Neonatal disorder
Is a type of
Legend
Disorder
Synonym
Body Structure
Relation

A concept in SNOMED CT




Disease
Disorder by body site
Infectious disease
Disorder of trunk
Infection by site
Bacterial infectious disease
Disorder of thorax
Respiratory tract infection
Bacterial infection by site
Disorder of lung
Lower respiratory tract infection
Lung consolidation
Bacterial respiratory infection
Pneumonia
Infectious disease of lung
Infective pneumonia
Bacterial lower respiratory infection
Bacterial pneumonia


Τ–20000
Αναπνευστικό σύστημα
Τ–25000
Τραχεία
Τ–28000
Πνεύμονας
Τ–28100
Δεξιός πνεύμονας
Τ–28500
Αριστερός πνεύμονας
Τ–28500
Βάση του αριστερού πνεύμονας
Τ–28530
Πύλη του αριστερού πνεύμονας

Snomed vs ICD

- ICD is a relatively ancient code family
  - Late 19th century roots
  - ICD9 was developed in 1970s! Even ICD10 is ~30 years old!
- ICD is a classification whereas SNOMED is a terminology
  - ICD tends to be more abstract. With SNOMED the user can get a more accurate description
  - ICD9 (or ICD10) tend to have a “unspecified” catch-all slot for most disorders.
- SNOMED is far more extensive than ICD
  - ICD only covers disorders
- SNOMED is implemented as an ontology
- Any number of relationships can be defined for each concept

UMLS

- a compendium of many controlled vocabularies in the biomedical sciences (created 1986)
- It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts.
- UMLS further provides facilities for natural language processing.
- It is intended to be used mainly by developers of systems in medical informatics.

UMLS

- Contains ~250.000 concepts and over 540.000 different terms.
- Integrates terms from
  - MeSh (273.000 terms), SNOMED (107.000), ICD–9–CM (35.000), DSM–IV (Διαγνωστικό και Στατιστικό Εγχειρίδιο Ψυχικών Διαταραχών), CPT (Τρέχουσα Ορολογία Επεμβατικών Διαδικασιών

UMLS Rest API

Base URI	Method Type	Path	Description
https://utslogin.nlm.nih.gov			
	POST	/cas/v1/tickets	Retrieves a Ticket Granting Ticket (TGT)
	POST	/cas/v1/tickets/{TGT}	Retrieves a single-use Service Ticket
GET	/search/{version}		Retrieves CUIs when searching by term or code
GET	/content/{version}/CUI/{CUI}		Retrieves information about a known CUI
GET	/content/{version}/CUI/{CUI}/atoms		Retrieves atoms and information about atoms for a known CUI
GET	/content/{version}/CUI/{CUI}/definitions		Retrieves definitions for a known CUI
GET	/content/{version}/CUI/{CUI}/relations		Retrieves NLM-asserted relationships for a known CUI
GET	/content/{version}/source/{source}/{id}		Retrieves information about a known source-asserted identifier
GET	/content/{version}/source/{source}/{id}/atoms		Retrieves information about atoms for a known source-asserted identifier
GET	/content/{version}/source/{source}/{id}/parents		Retrieves immediate parents of a source-asserted identifier
GET	/content/{version}/source/{source}/{id}/children		Retrieves immediate children of a source-asserted identifier
GET	/content/{version}/source/{source}/{id}/ancestors		Retrieves all ancestors of a source-asserted identifier
GET	/content/{version}/source/{source}/{id}/descendants		Retrieves all descendants of a source-asserted identifier
GET	/content/{version}/source/{source}/{id}/relations		Retrieves all relationships of a source-asserted identifier
GET	/subsets/{version}/		Retrieves all available source-asserted subsets
GET	/subsets/{version}/source/{source}/{id}		Retrieves information about a specific source-asserted subset
GET	/subsets/{version}/source/{source}/{id}/members		Retrieves members of a specific source-asserted subset
GET	/subsets/{version}/source/{source}/{id}/member/{id}		Retrieves an individual member of a specific source-asserted subset
GET	/content/{version}/source/{source}/{id}/attributes		Retrieves information about source-asserted attributes
GET	/semantic-network/{version}/TUI/{id}		Retrieves information for a known Semantic Type identifier (TUI)
GET	/content-views/{version}		Retrieve all available content views in the UMLS
GET	/content-views/{version}/CUI/{CUI}		Retrieve information about a specific content view
GET	/content-views/{version}/CUI/{CUI}/members		Retrieve members of a specific content view
GET	/content-views/{version}/CUI/{CUI}/member/{id}		Retrieves an individual member of a specific content view

Natural Language Processing

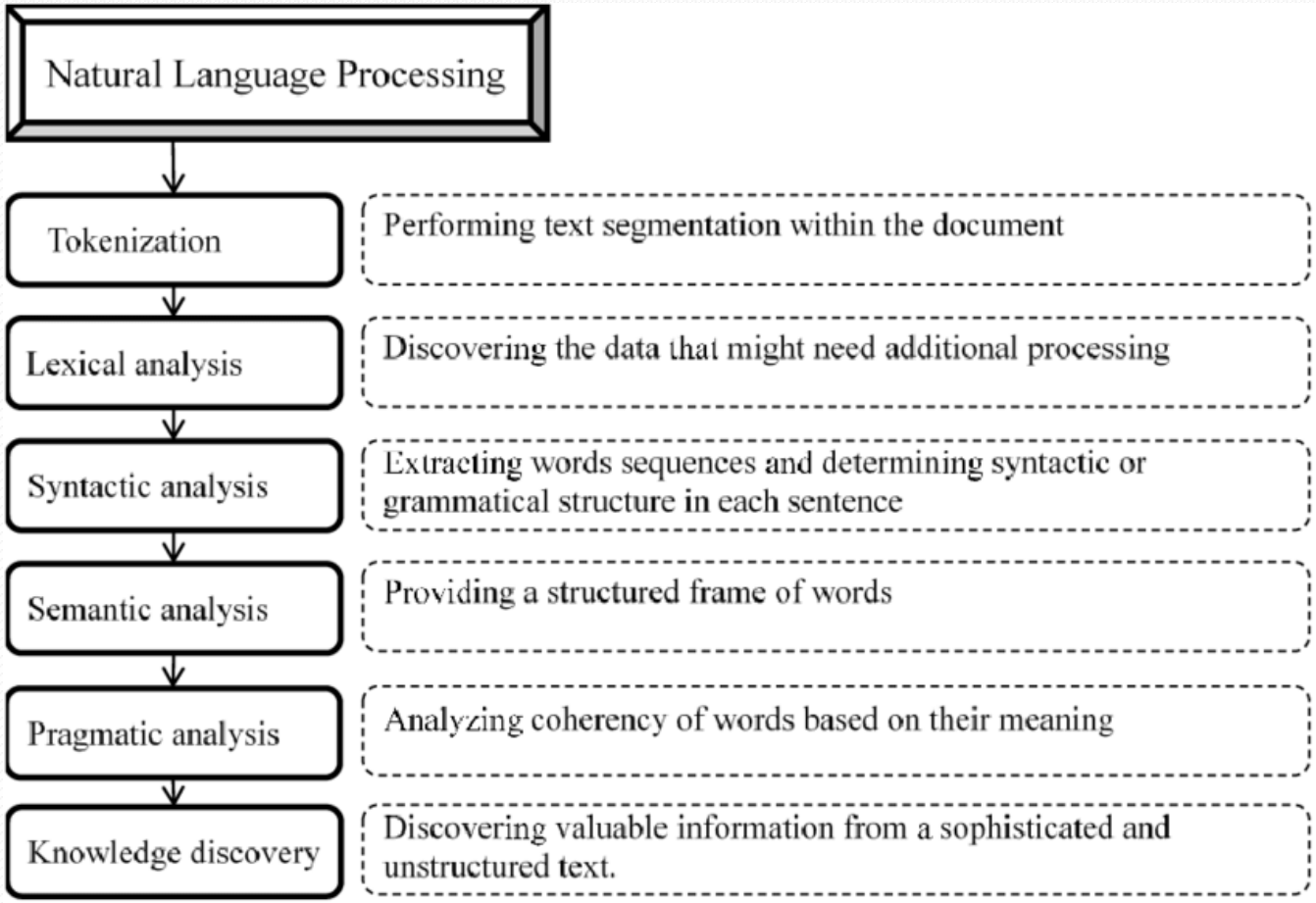
NLP

- Natural Language Processing
  - From Text to Meaning

- Combines:
  - Computer Science
  - Linguistics
  - Machine Learning

Goal: Convert **text → structured meaning**

NLP steps


Natural Language Processing
Tokenization
Performing text segmentation within the document
Lexical analysis
Discovering the data that might need additional processing
Syntactic analysis
Extracting words sequences and determining syntactic or grammatical structure in each sentence
Semantic analysis
Providing a structured frame of words
Pragmatic analysis
Analyzing coherency of words based on their meaning
Knowledge discovery
Discovering valuable information from a sophisticated and unstructured text.

NLP common patterns

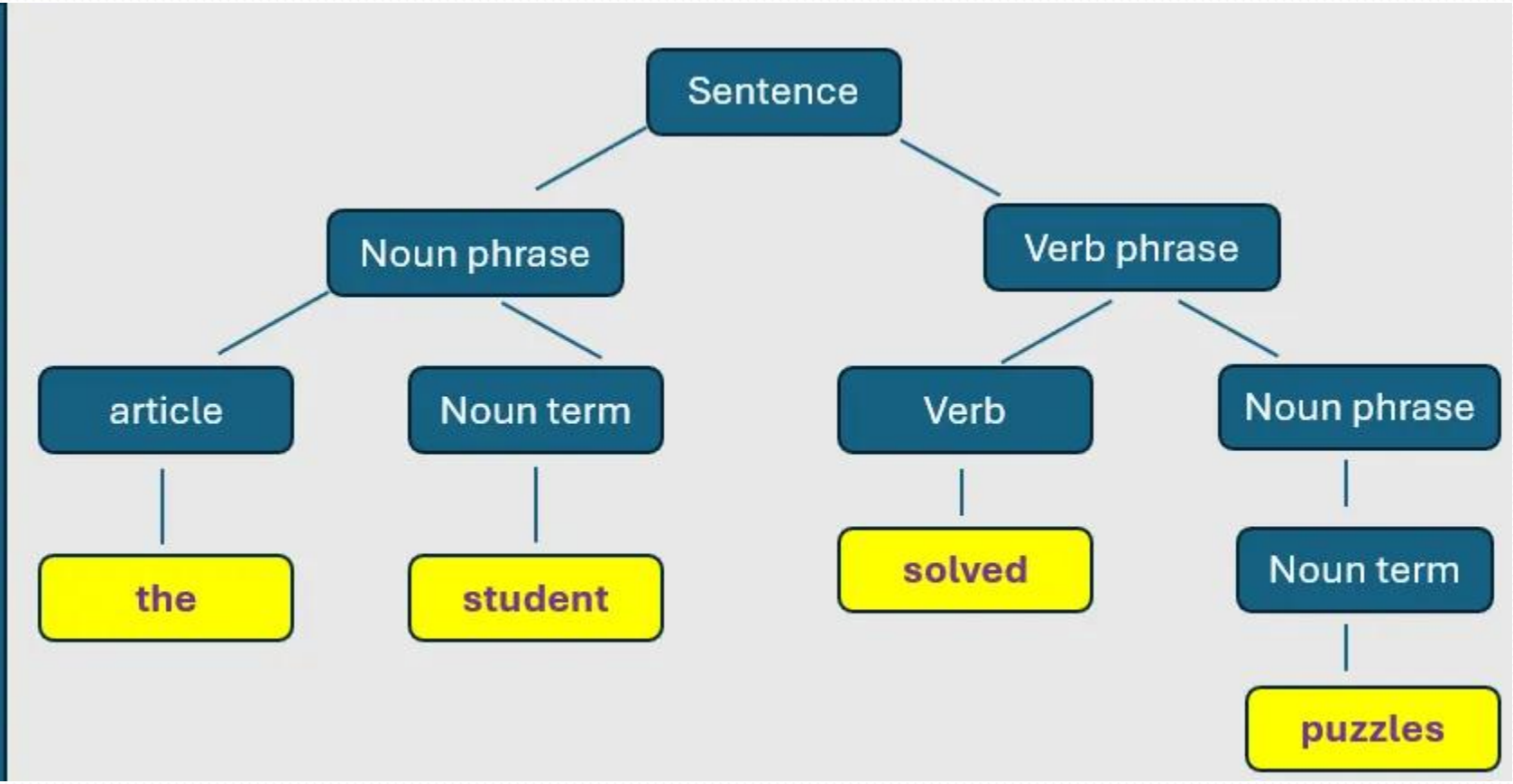
The syntactic analysis breaks sentences into parts to understand their structure. Common patterns include:

- **Subject-Verb-Object (SVO):** A basic sentence structure where the subject does something to the object. e.g., “The bird (subject) catches (verb) the worm (object).”
- **Complex Sentences:** Sentences with more than one part, needing more work to understand. e.g., “After the game ended, she went to bed” has two parts: one about the game and one about going to bed.

Syntactic Analysis

The grammar rules will look like these:

Sentence -> NP VP
NP -> Article Noun
NP-> Noun
VP -> Verb NP
Noun -> student | NN
Noun -> puzzles | NNS
Article -> the
Verb -> solved | VBD



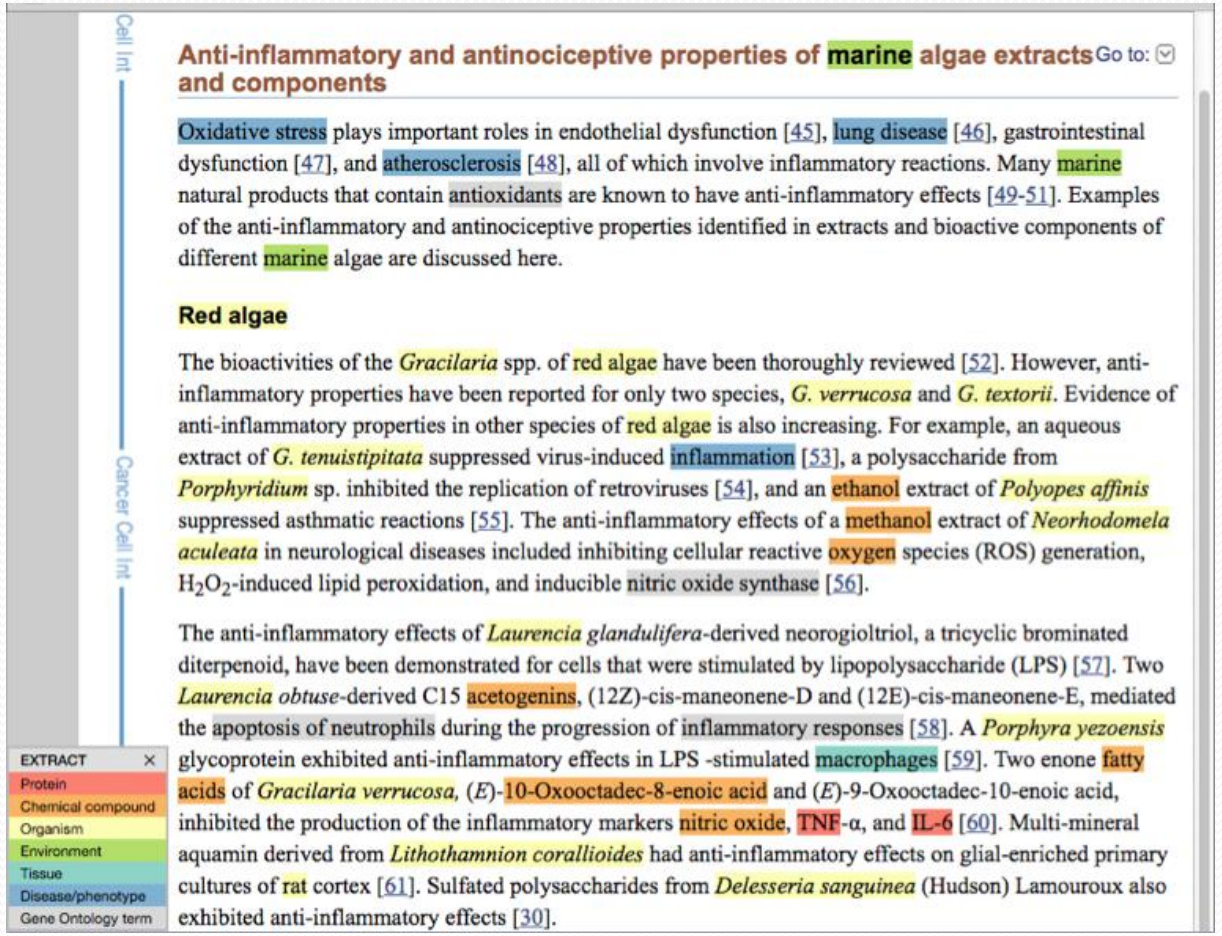
NLP example

EXTRACT: Interactive Extraction of Metadata

- https://extract.jensenlab.org

Identifies

- genes/proteins,
- chemical compounds,
- organisms,
- environments,
- tissues,
- diseases,
- Phenotypes
- Gene Ontology terms



Example
calchas.ics.forth.gr

Problem definition

Bioinformatics community handle an increasingly large and diverse set of tools.

- Many accessible biomedical resources and tools repositories exist and expand
  - ***Taverna – myExperiment, Galaxy, BioCatalogue, SeqAnswers, Bioconductor, Elixir (bio.tools)***

Search for various biomedical tools for use **by researchers**

Tools to search:

- generic Google-like searches
- specialized web forums to seek technical and analytical advices

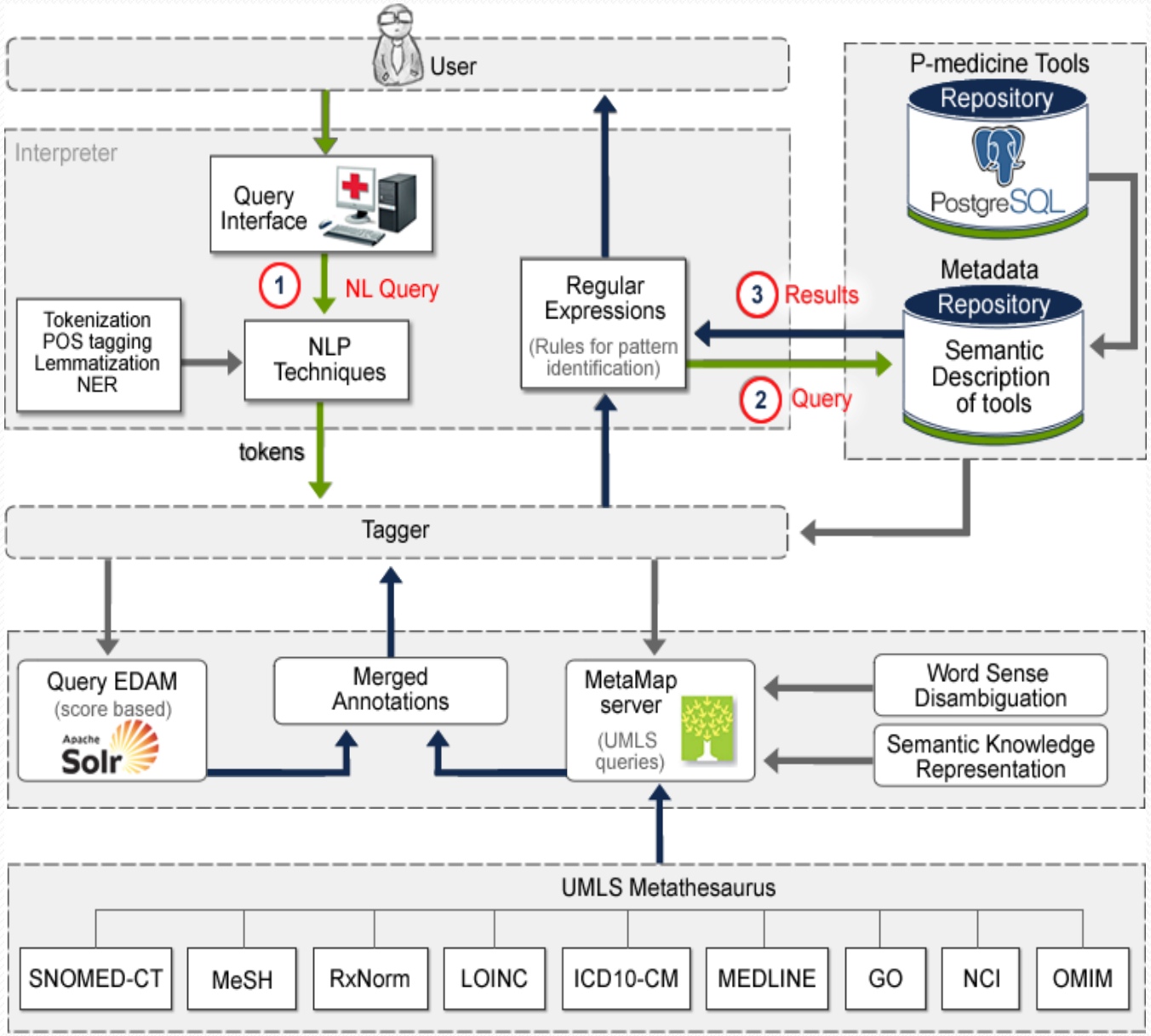
- Thousands of tools to search in more than one location
  - **Time consuming**
- Information about resources require IT knowledge
  - **End users in most of the cases are not IT experts**

Researchers demand ever more powerful and convenient means to organize, find, understand, compare, select, use and connect the available resources.

Proposed Solution

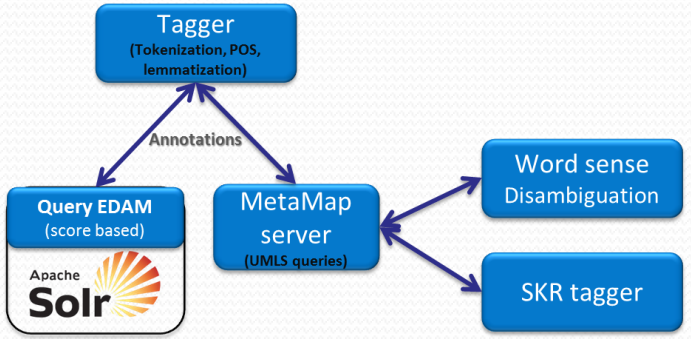
Dynamic service discovery system on natural language text:

- Researcher's request in natural language
  - Syntactic analysis of the request with NLP techniques
    - need for **Natural language processing (NLP)**
  - Sentence is semantically annotated with ontologies
    - need for **Semantic Annotation & Named-entity recognition (NER)**
- Tools repository
  - annotated tools with ontologies
    - need for **Ontologies**
- Match research questions with the candidate tools


User
Interpreter
Query Interface
1
NL Query
Tokenization
POS tagging
Lemmatization
NER
NLP Techniques
tokens
Regular Expressions
(Rules for pattern identification)
3
Results
2
Query
P-medicine Tools
Repository
PostgreSQL
Metadata
Repository
Semantic Description of tools
Tagger
Query EDAM
(score based)
Apache Solr
Merged Annotations
MetaMap server
(UMLS queries)
Word Sense Disambiguation
Semantic Knowledge Representation
UMLS Metathesaurus
SNOMED-CT
MeSH
RxNorm
LOINC
ICD10-CM
MEDLINE
GO
NCI
OMIM

The tagger

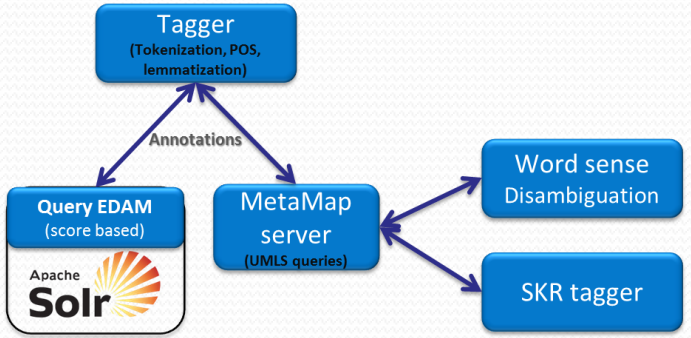
MetaMap Server



A UMLS toolkit

- unit of files and software – 160 controlled vocabularies
- 11,2 million terms – 2,8 million concepts
- 9 ontologies were chosen:
  - MeSH, RxNorm, SNOMED CT, Loinc, ICD10-CM, Medline, NCI, OMIM, GO
- Semantic Knowledge Representation
  - a stochastic part of speech tagger employing a hidden Markov model (HMM) to combine contextual information with lexical information to improve on baseline tagging accuracy
- Word Sense Disambiguation server
  - E.g. the word "cold" has several senses and may refer to a disease, a temperature sensation, or an environmental condition

The tagger EDAM



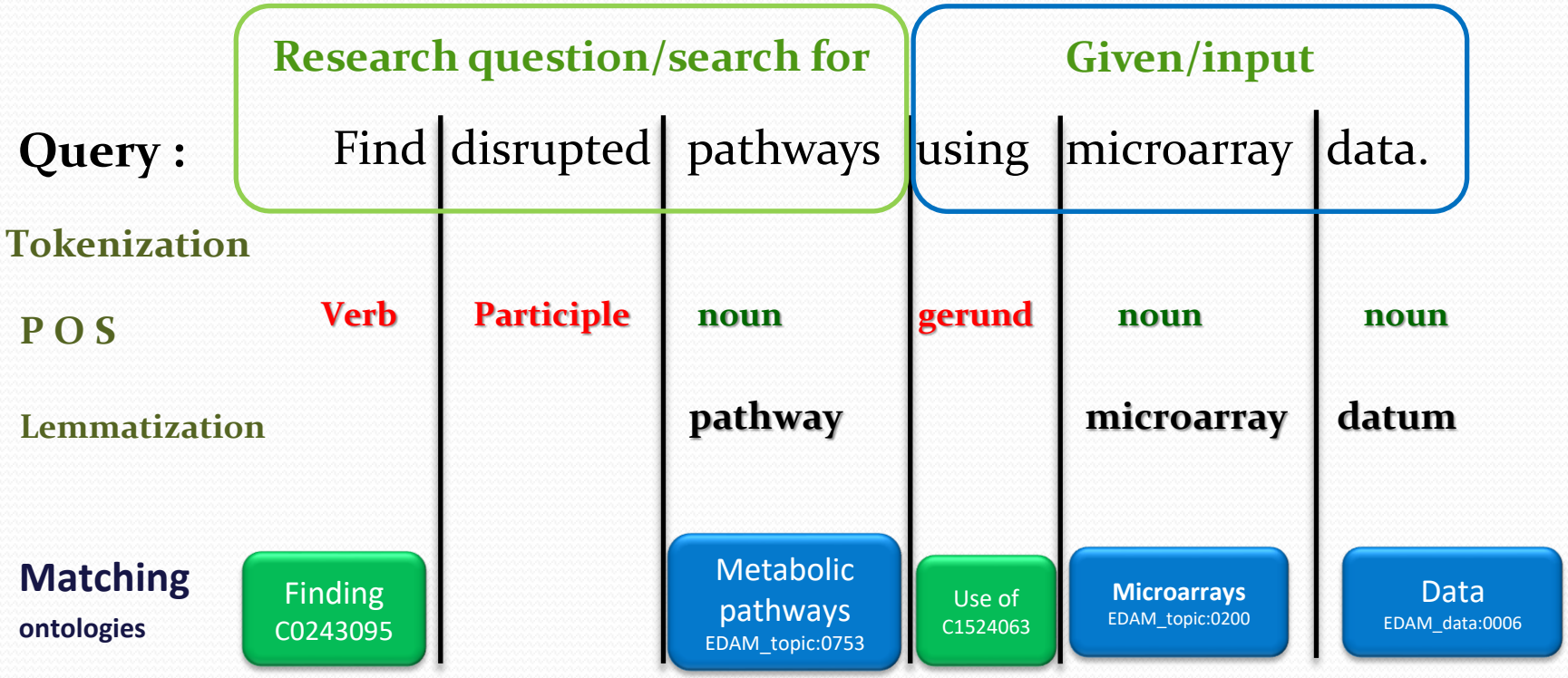
- EDAM (EMBRACE Data and Methods)
  - includes over 2200 defined common bioinformatics operations concepts
- We created an EDAM tagger using Apache Solr full text server


```
[
{
        "id": "EDAM_format:3008",
        "name": "MAF",
        "subset": [
                "bioinformatics",
                "edam",
                "formats"
        ]
,
        "created_in": "beta12orEarlier",
        "def": "Multiple Alignment Format (MAF) supporting alignments of whole genomes with rearra
ngements, directions, multiple pieces to the alignment, and so forth. [http://edamontology.org]",
        "namespace": "format",
        "is_a": [
                "EDAM_format:2330 ! Textual format",
                "EDAM_format:2554 ! Alignment format (text)",
                "EDAM_format:2919 ! Sequence annotation track format"
        ]
,
        "xref": "format5",
        "comment": "Typically generated by Multiz and TBA aligners; can be displayed in a genome b
rowser like a sequence annotation track. This should not be confused with MIRA Assembly Format or
Mutation Annotation Format."}
]
```


Score based – Weight formula:

$$(id*10) + (name*10) + (synonym*6) + (subset*3) + (is\_a*3) + (def*2) + (comment*1)$$

An example


Research question/search for
Given/input
Query :
Find
disrupted
pathways
using
microarray
data.
Tokenization
P O S
Verb
Participle
noun
gerund
noun
noun
Lemmatization
pathway
microarray
datum
Matching
ontologies
Finding
C0243095
Metabolic pathways
EDAM_topic:0753
Use of
C1524063
Microarrays
EDAM_topic:0200
Data
EDAM_data:0006

Tools Repository

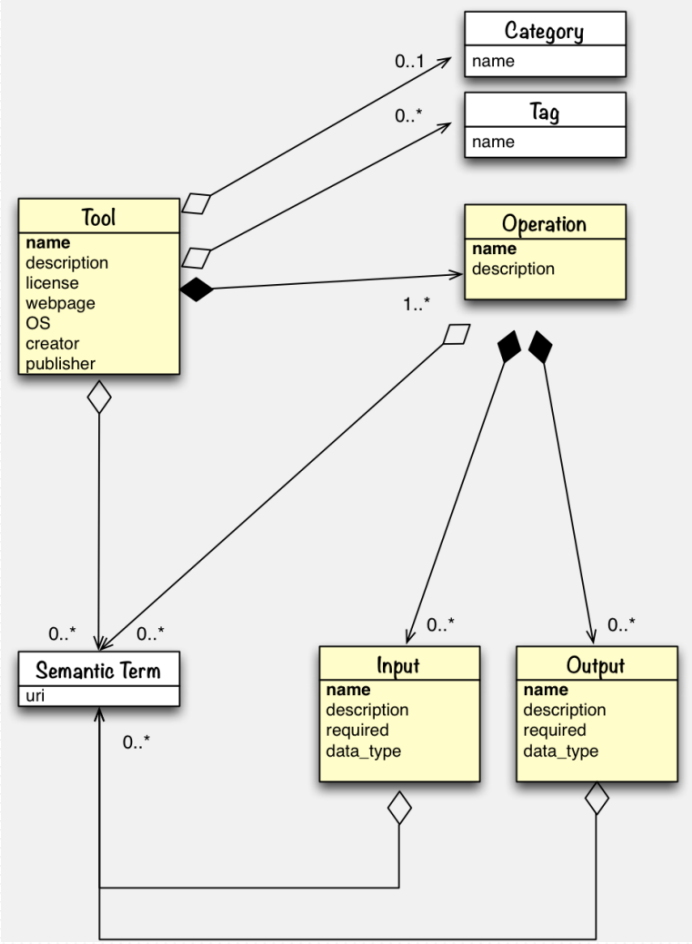
Tools coming form Elixir bioinformatics tools repository (https://bio.tools)

- About 6200 tools registered

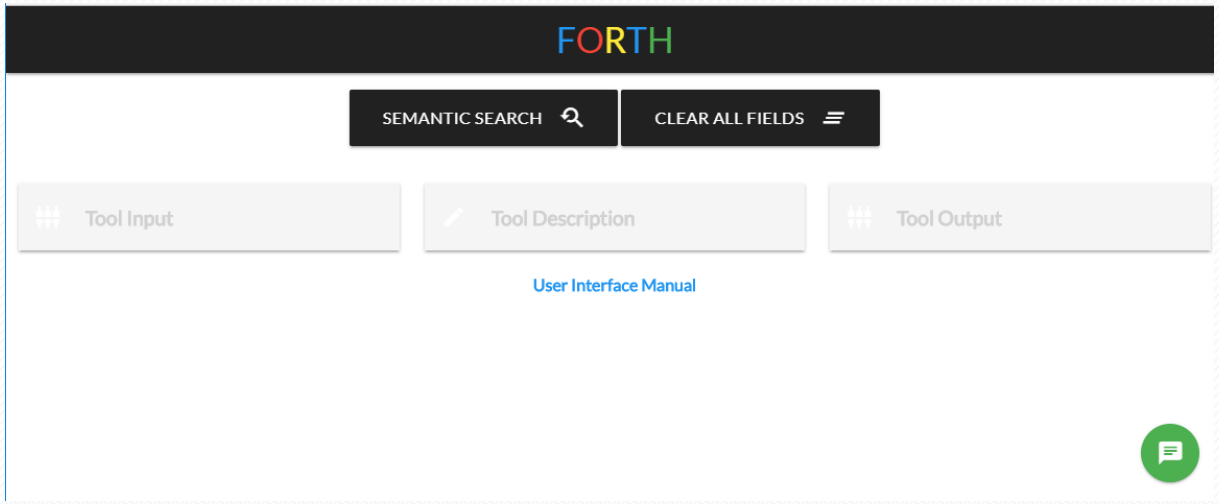
  Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., Anthon, C., Beard, N., Berka, K., Bolser, D. and Booth, T., 2015. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic acids research*, p.gkv1116.

Storage

- Auto annotation
- Auto updates (from bio.tools)
- Specific schema
  - Full text search capabilities



User interface (input)



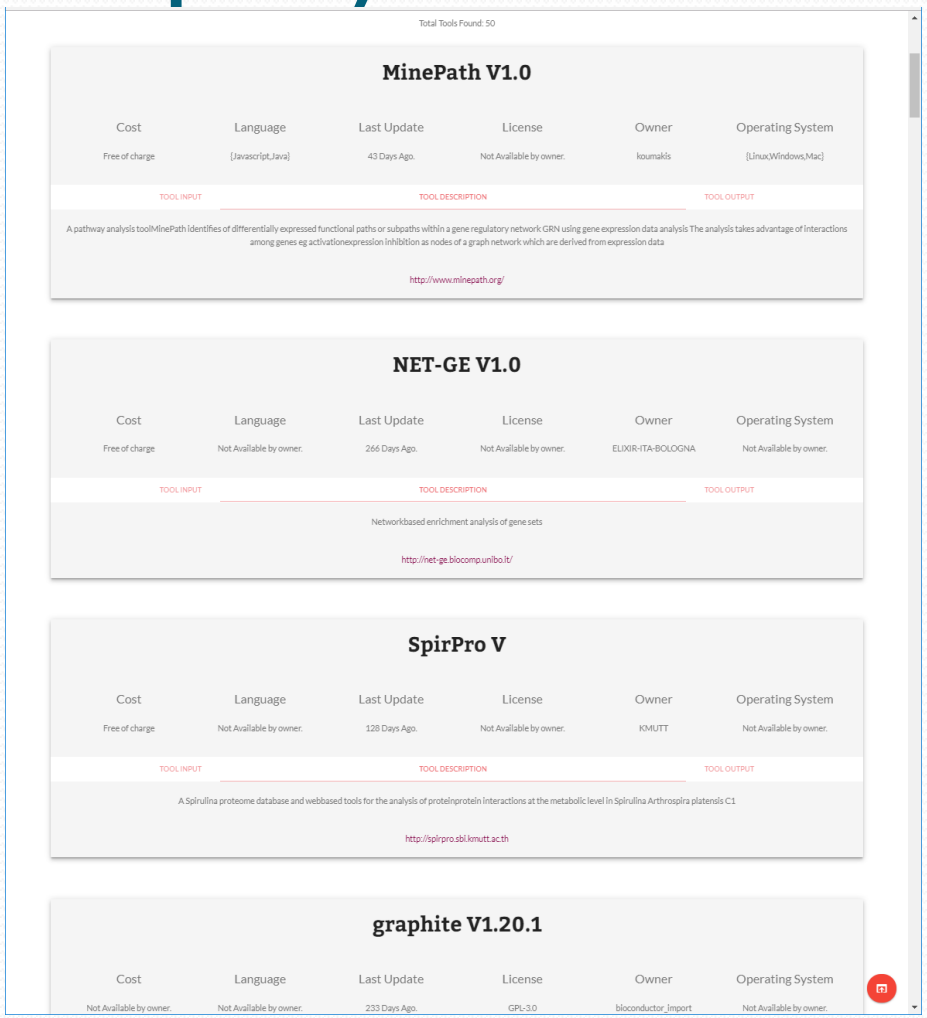
- *Tool Input*: you can enter the input data that you want the tool to be able to support.
- *Tool Description*: match with a tool's description or tool's name your are searching for.
- *Tool Output*: you can enter the output data that you want the tool to be able to support.

User Interface (output)

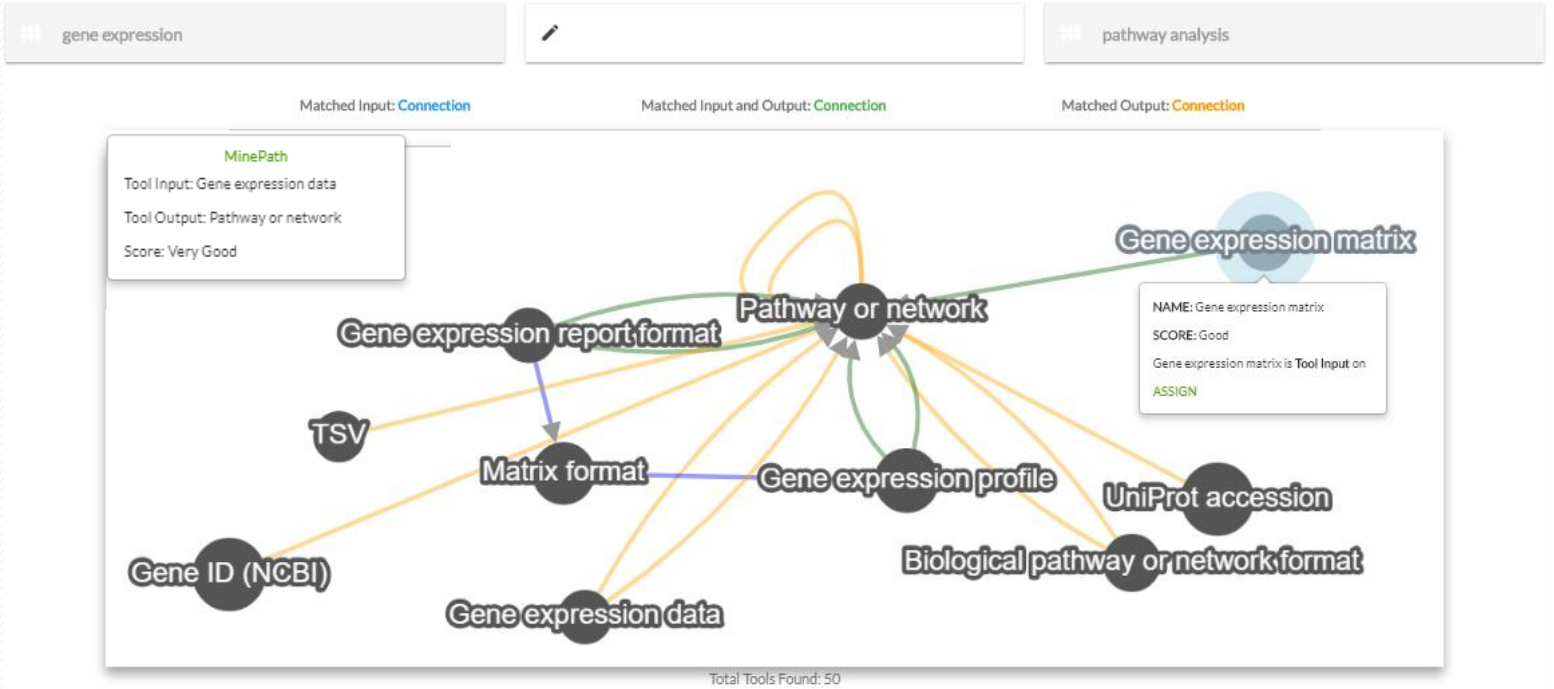
Tool Cards

Each tool card provides the following:

- Tool's name.
- Tool's input.
- Tool's description.
- Tool's output.
- Tool's cost.
- Tool's source language.
- Tool's last update.
- Tool's license.
- Tool's owner.
- Tool's operating System.
- Tool's homepage (URL).



User Interface (output)



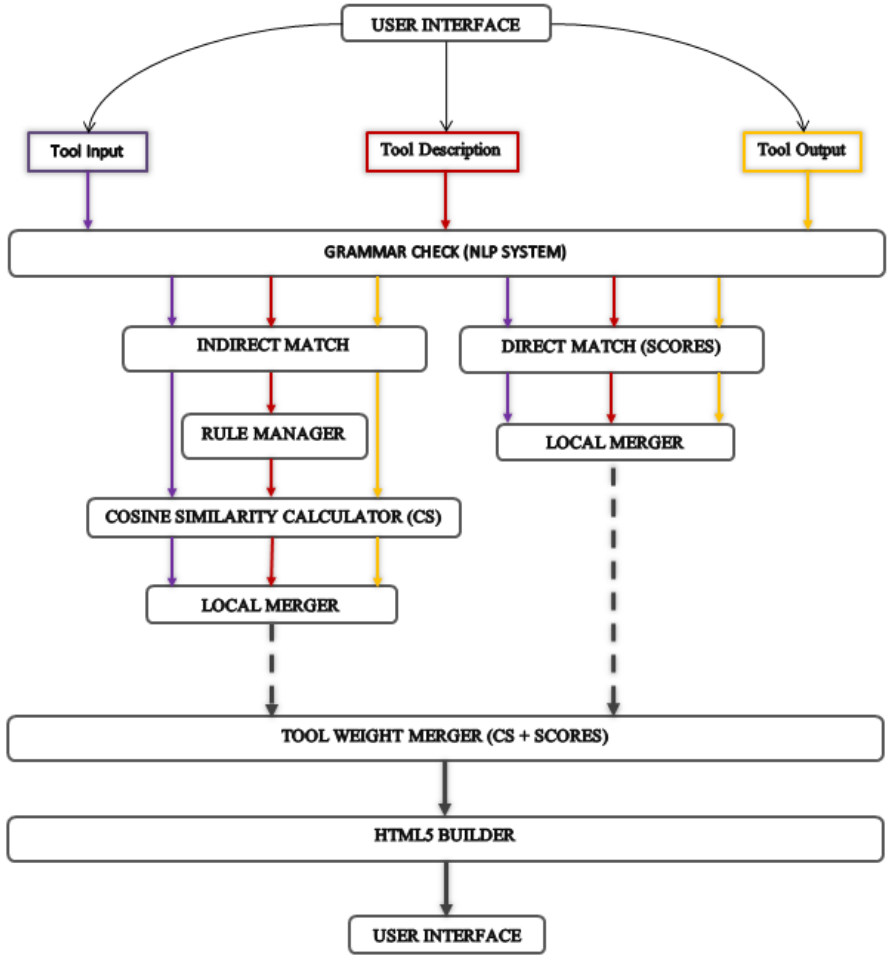
Graph

Nodes represent data format

Edges represent tools

- **Blue edges**: Your requested input data/concept exist for this tool
- **Orange edges**: Your requested output data/concept exist for this tool
- **Green edges**: Your requested input and output data/concept exist for this tool.

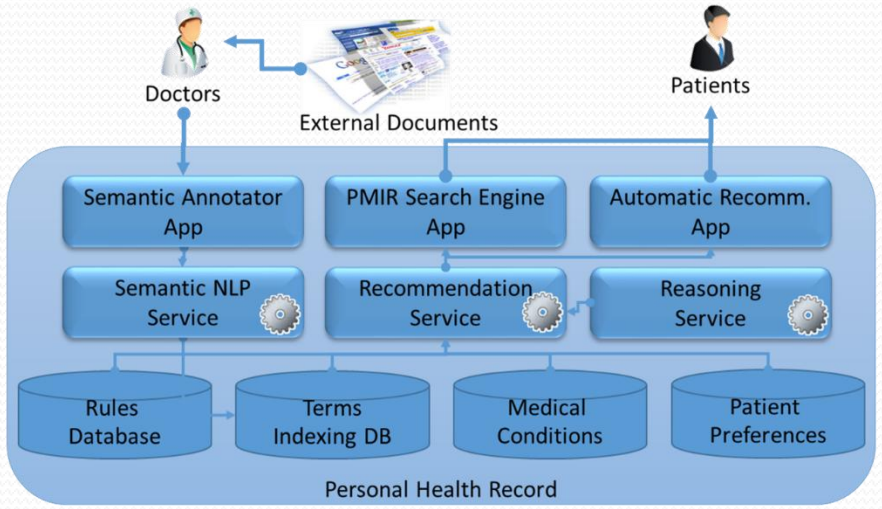
Calchas : Flow of Operations



Example
Personal Health
Information Recommender

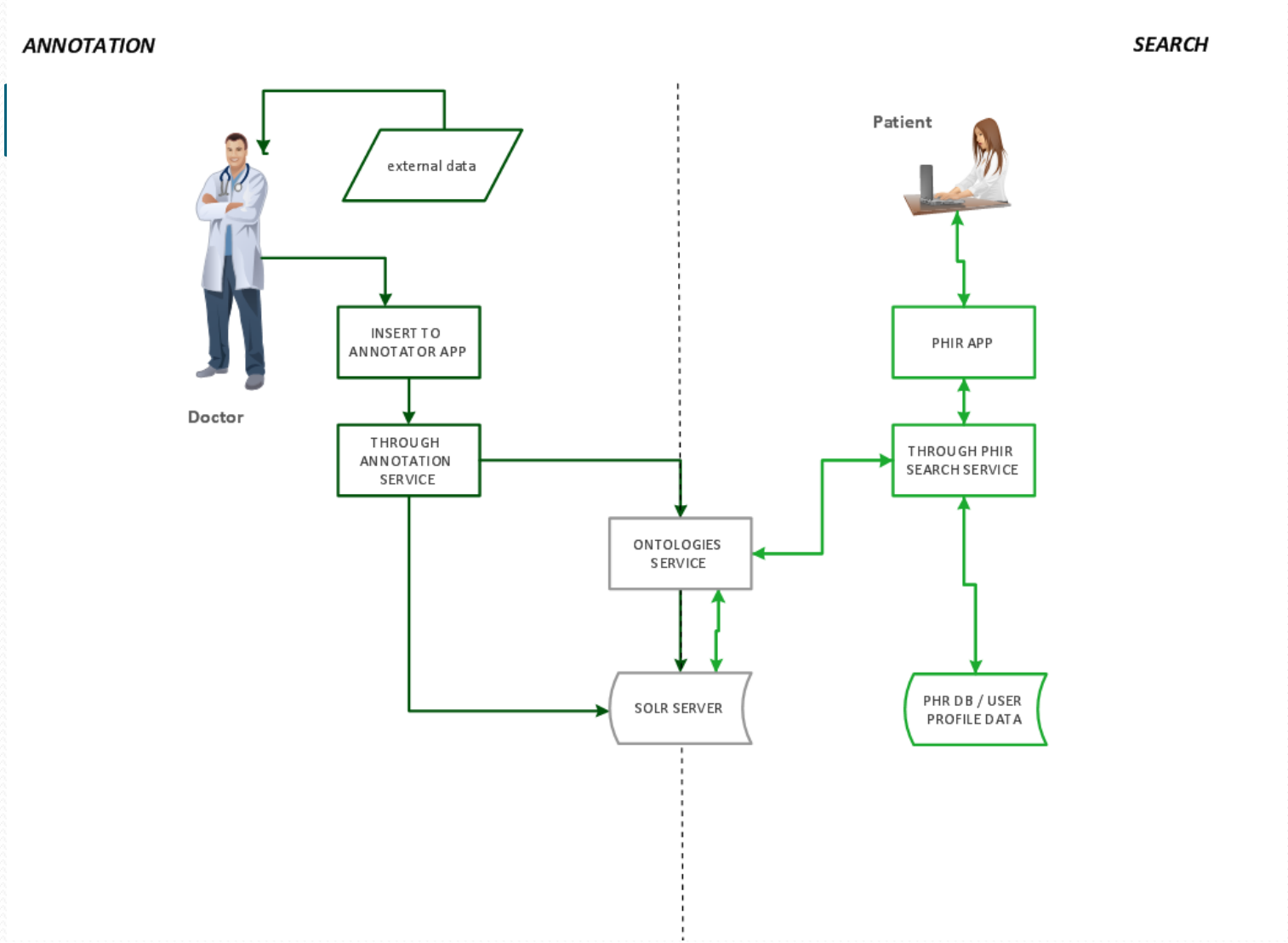
Personal Health Information Recommender (PHIR)

- ✓ allows searching in a high quality document repository
- ✓ automatically provides intelligent and personalized recommendations, according to the individual preferences and medical conditions

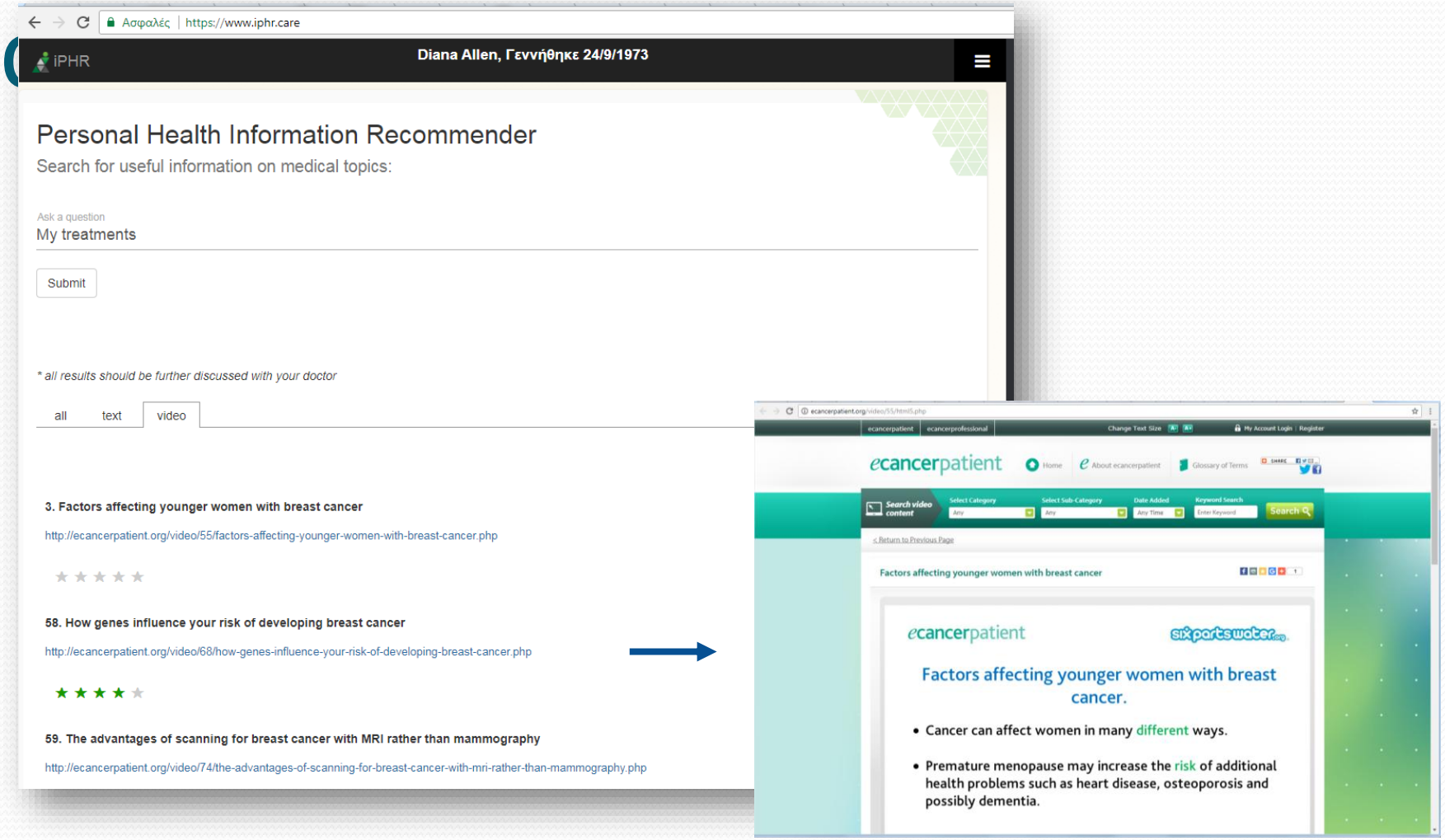


Integrated into a PHR as a set of individual apps

PHIR a

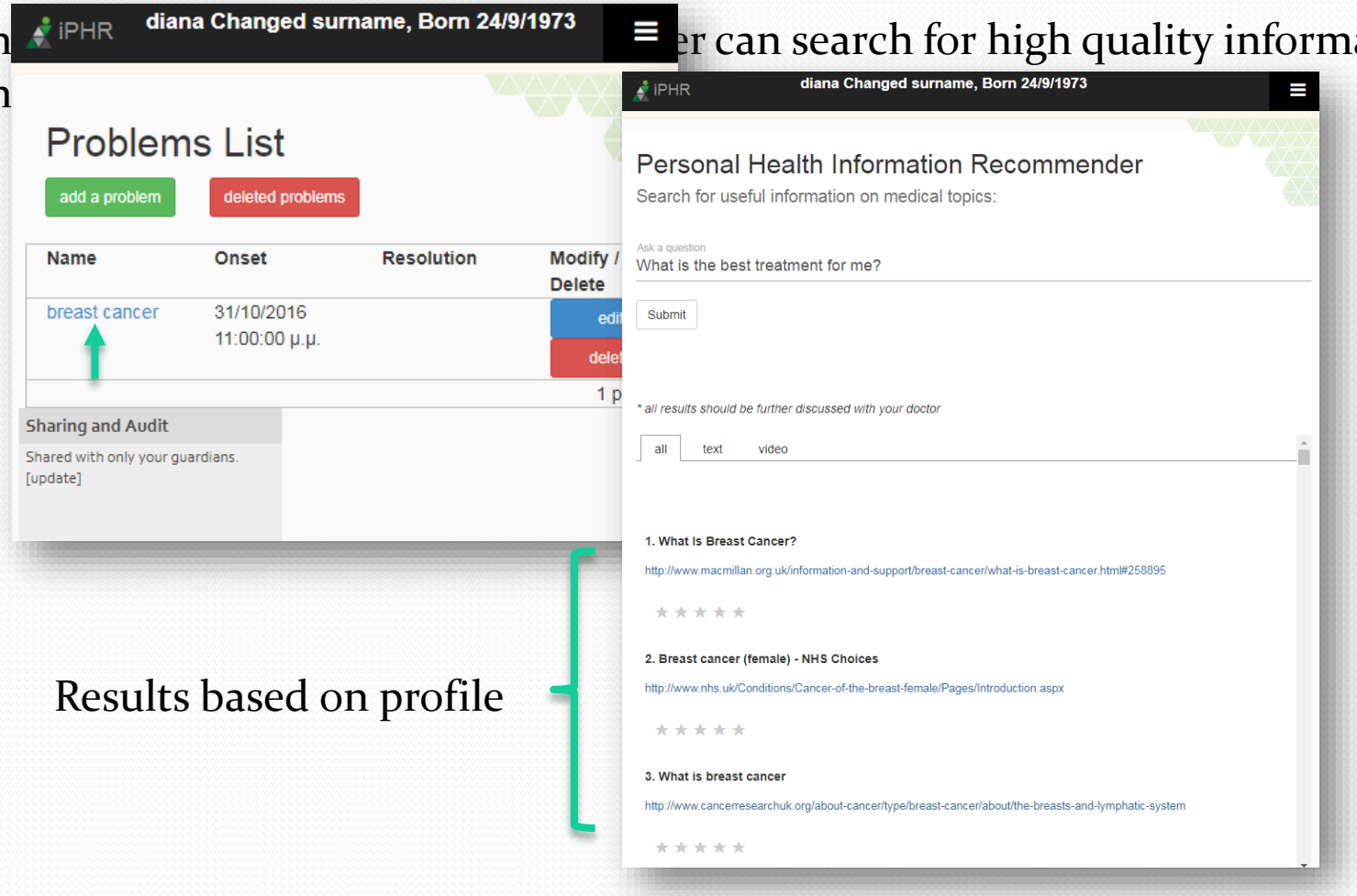

ANNOTATION
SEARCH
external data
Patient
INSERT TO ANNOTATOR APP
PHIR APP
Doctor
THROUGH ANNOTATION SERVICE
THROUGH PHIR SEARCH SERVICE
ONTOLOGIES SERVICE
SOLR SERVER
PHR DB / USER PROFILE DATA

Searc


Ασφαλές | https://www.iphr.care
iPHR
Diana Allen, Γεννήθηκε 24/9/1973
Personal Health Information Recommender
Search for useful information on medical topics:
Ask a question
My treatments
Submit
* all results should be further discussed with your doctor
all
text
video
3. Factors affecting younger women with breast cancer
http://ecancerpatient.org/video/55/factors-affecting-younger-women-with-breast-cancer.php
58. How genes influence your risk of developing breast cancer
http://ecancerpatient.org/video/68/how-genes-influence-your-risk-of-developing-breast-cancer.php
59. The advantages of scanning for breast cancer with MRI rather than mammography
http://ecancerpatient.org/video/74/the-advantages-of-scanning-for-breast-cancer-with-mri-rather-than-mammography.php
ecancerpatient
Factors affecting younger women with breast cancer.
Cancer can affect women in many different ways.
Premature menopause may increase the risk of additional health problems such as heart disease, osteoporosis and possibly dementia.

Search Engine App & Service

Th er can search for high quality information.
Th



Results based on profile