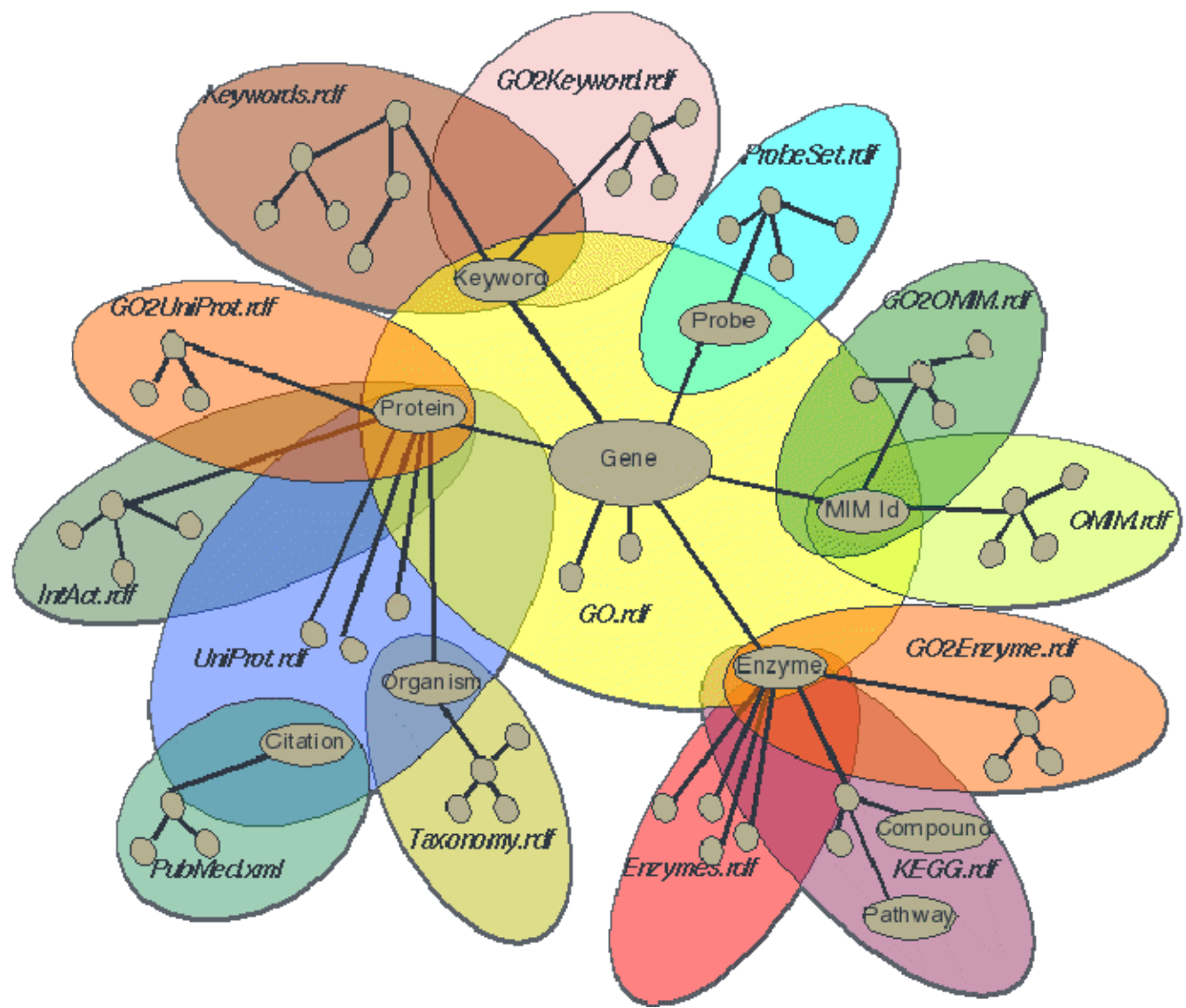


Ontologies

examples and applications

Steffen Staab

Semantic Web



- Framework consisting of several knowledge bases and according tools
- Goal
 - ◆ Improvement of knowledge management in information systems with medical context
- Publisher
 - ◆ US National Library of Medicine



<http://www.nlm.nih.gov/research/umls/>

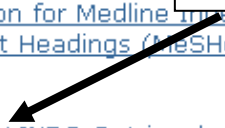
- Automated processing of biomedical and health care related terms
- Support of application developers
- Integration of several kinds of information, e.g.:
 - ◆ Scientific literature
 - ◆ Medical records
 - ◆ Epidemiological data

Databases

- [AIDS Information Resources](#)
- [Authorship in MEDLINE](#)
- [Chemical Carcinogenesis Research Information System \(CCRIS®\)](#)
- [ChemIDplus](#)
- [ClinicalTrials.gov](#)
- [Conflict of Interest Disclosure and Journal Supplements in MEDLINE®](#)
- [Construction of National Library of Medicine Title Abbreviations](#)
- [Developmental and Reproductive Toxicology/Environmental Teratology Information Center \(DART®/ETIC\) Database](#)
- [DIRLINE®](#)
- [DOCLINE®](#)
- [Entrez Molecular Sequence Database System](#)
- [GenBank Retrieval and Sequence Similarity Searching by Internet Electronic Mail](#)
- [GENE-TOX](#)
- [Haz-Map](#)
- [Hazardous Substances Data Bank \(HSDB®\)](#)
- [Health Services Research Programs](#)
- [Health Services Technology](#)
- [Household Products Database](#)
- [Human Genome Resources](#)
- [Images from the History of Medicine](#)
- [Integrated Risk Information System](#)
- [International Medlars Centers](#)
- [Journal Selection for MEDLINE®](#)
- [Journal Selection for Medline Indexing at NLM, FAQ](#)
- [Medical Subject Headings \(MeSH®\)](#)
- [MEDLINE®](#)
- [NLM Gateway](#)
- [PubMed®: MEDLINE® Retrieval on the World Wide Web](#)
- [SERHOLD®](#)
- [Submitting Data to GenBank®](#)
- [Tox Town](#)
- [Toxic Chemical Release Inventory \(TRI®\)](#)
- [TOXLINE®](#)
- [TOXMAP: Environmental Health e-Maps](#)
- [TOXNET: Toxicology Data Network](#)
- [Visible Human Project®](#)
- [WISER](#)

PubMed

Over 5,000 biomedical journals
Coverage extends back to 1950s
citations annotated with **Medical Subject Headings (MeSH)**



Databases

- [AIDS Information Resources](#)
- [Authorship in MEDLINE](#)
- [Chemical Abstracts](#)
- [ChemID](#)
- [ClinicalTrials.gov](#)
- [Conflict of Interest](#)
- [Construction of the UMLS](#)
- [Development of the UMLS](#)
- [DIRLINE](#)
- [DOCLINE](#)
- [Entrez](#)
- [GenBank](#)
- [GENE-T](#)
- [Haz-Mat](#)
- [Hazardous Waste](#)
- [Health](#)
- [Health](#)
- [Household](#)
- [Human](#)
- [Images](#)
- [Integrating](#)
- [International Medlars Centers](#)
- [Journal Selection for MEDLINE](#)
- [Journal Selection for Medline Indexing at NLM, FAQ](#)
- [Medical Subject Headings \(MeSH\)](#)
- [MEDLINE](#)
- [NLM Gateway](#)
- [PubMed: MEDLINE Retrieval on the World Wide Web](#)
- [SERHOLD](#)
- [Submitting Data to GenBank](#)
- [Tox Town](#)
- [Toxic Chemical Release Inventory \(TRI\)](#)
- [TOXLINE](#)
- [TOXMAP: Environmental Health e-Maps](#)
- [TOXNET: Toxicology Data Network](#)
- [Visible Human Project](#)
- [WISER](#)

Internet Services

- [ChemIDplus](#)
- [DOCLINE](#)
- [Electronic Document Delivery, Usage and Conversion: DocView, DocMorph and MyMorph](#)
- [Images from the History of Medicine \(IHM\)](#)
- [LocatorPlus](#)
- [MedlinePlus](#)
- [Profiles in Science](#)
- [PubMed: MEDLINE](#)
- [TOXMAP: Environmental Health e-Maps](#)
- [ToxMystery](#)
- [TOXNET: Toxicology Data Network](#)
- [ToxSeek](#)
- [UMLS Knowledge](#)
- [WISER](#)

LocatorPlus

Online catalog of National Library of Medicine (NLM) over 1.2 million catalog records (books, journals, multimedia, ...)

Databases

- [AIDS Information Resources](#)
- [Authorship in MEDLINE](#)
- [Chemical Abstracts](#)
- [ChemIDplus](#)
- [ClinicalTrials.gov](#)
- [Conflict of Interest](#)
- [Construction of the UMLS](#)
- [Development of the UMLS](#)
- [DIRLINE](#)
- [DOCLINE](#)
- [Entrez](#)
- [GenBank](#)
- [GENE-T](#)
- [Haz-Mat](#)
- [Hazardous Waste](#)
- [Health](#)
- [Health](#)
- [Household](#)
- [Human](#)
- [Images](#)
- [Integrating](#)
- [International Media](#)
- [Journal Selection for](#)
- [Journal Selection for](#)
- [Medical Subject Headings](#)
- [MEDLINE](#)
- [NLM Gateway](#)
- [PubMed: MEDLINE](#)
- [SERHOLD](#)
- [Submitting Data to](#)
- [Tox Town](#)
- [Toxic Chemical Release Inventory \(TRI\)](#)
- [TOXLINE](#)
- [TOXMAP: Environmental Health e-Maps](#)
- [TOXNET: Toxicology Data Network](#)
- [Visible Human Project](#)
- [WISER](#)

Internet Services

- [ChemIDplus](#)
- [DOCLINE](#)
- [Electronic Document Delivery, Usage and Conversion: DocView, DocMorph and MyMorph](#)
- [Images from the History of Medicine](#)
- [LocatorPlus](#)
- [MedlinePlus](#)
- [Profiles in Science](#)
- [PubMed: MEDLINE Retrieval](#)
- [TOXMAP: Environmental Health e-Maps](#)
- [ToxMystery](#)
- [TOXNET](#)
- [ToxS](#)
- [UMLS](#)
- [WISER](#)

Collection and Indices

- [Access to Audiovisual Materials](#)
- [Errata, Retractions, Duplicates](#)
- [FAQ: Journal Selection for](#)
- [IndexCat™](#)
- [Interlibrary Loan](#)
- [Loansome Doc - A Document Ordering Feature of PubMed and the NLM Gateway](#)
- [Medical Subject Headings \(MeSH\)](#)
- [NLM Classification](#)
- [Online Indexing System](#)
- [Preservation Program](#)
- [Reference and Customer Services](#)

IndexCat
3.7 million references (5th – 20th century !)
Journal articles - 2,500,000 references
Dissertations and theses - 470,000 references
Monographs - 616,000 references
Journal titles - 32,000 references
Portraits - 4,000 references

- Linked knowledge bases
 - ◆ Metathesaurus
 - ◆ Semantic Network
 - ◆ SPECIALIST Lexicon & NLP system

- Tools
 - ◆ UMLS Knowledge Source Server
 - ◆ MetamorphoSys
 - ◆ Lexical Tools

- Data stock
 - ◆ Terms related to medicine and health care
 - ◆ Additional detailed information
 - ◆ Integration of heterogeneous thesauri and classifications

- Goals
 - ◆ Linkage between different terms/views of the same concept
 - ◆ Generation of relations between different concepts

- It must be *customized* for efficient use ...

- Alcohol and Other Drug Thesaurus
- Physicians' Current Procedural Terminology
- International Classification of Diseases (ICD)
- Gene Ontology (GO)
- Health Level Seven Vocabulary (HL7)
- MEDLINE
- Medical Subject Headings (MESH)
- Systematized Nomenclature of Medicine (SNOMED)
- ...

→ More than 100 vocabularies – more than 1 million concepts

- Elements with unique identifier
 - ◆ Concepts (CUI)
 - ◆ Strings: different languages/spellings, etc. (SUI)
 - ◆ Atoms: every appearance of a string (AUI)
 - ◆ Terms: Groups lexically related strings (TUI)
- Relations between concepts
- Attributes for concepts, atoms und relations

Contradictory representations are adopted
→ No overall, consistent ontology

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs)
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	L0004238 Atrial Fibrillation (preferred) Atrial Fibrillations	S0016668 Atrial Fibrillation (preferred)	A0027665 Atrial Fibrillation (from MSH)
			A0027667 Atrial Fibrillation (from PSY)
		S0016669 (plural variant) Atrial Fibrillations	A0027668 Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	S0016899 Auricular Fibrillation (preferred)	A0027930 Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	A0027932 Auricular Fibrillations (from MSH)

- Source internal relations
 - ◆ Within one source vocabulary
 - ◆ Indicated e.g. by hierarchies, cross references
 - ◆ Statistical Relations, e.g. joint appearance of concepts in articles, several diagnose codes for the same patient, etc.

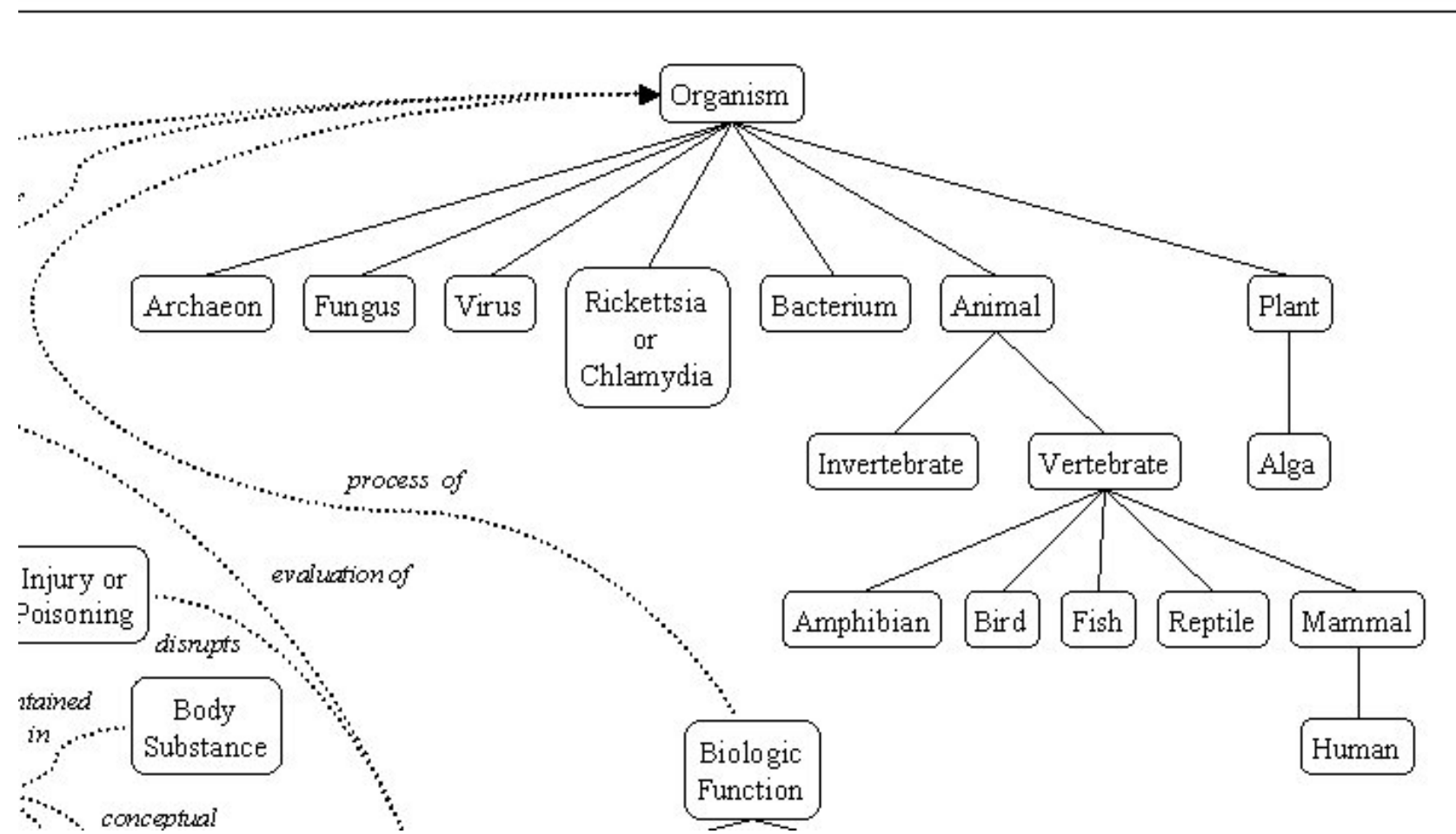
- Source overlapping relations
 - ◆ Mappings between multiple source vocabularies
 - ◆ Inclusion of „orphaned“ concepts into more detailed contexts of other vocabularies

- Goal
 - ◆ Consistent classification of concepts from the Metathesaurus

- Data stock:
 - ◆ Topic categories: „Semantic Types“
 - ◆ Relations between Semantic Types: „Semantic Relations“
 - ◆ 135 Semantic Types, 54 Semantic Relations

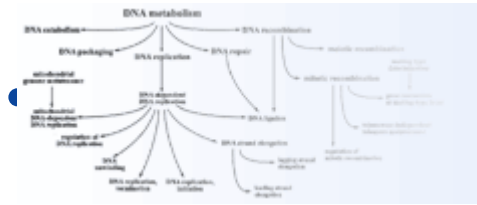
- Each concept in the Metathesaurus is associated to at least one Semantic Type

- Main concepts for Semantic Types
 - ◆ Organisms
 - ◆ Anatomical structures
 - ◆ Biologic function
 - ◆ Chemicals
 - ◆ Events
 - ◆ Physical objects
 - ◆ Concepts or ideas
- Main concepts for Semantic Relations
 - ◆ is a
 - ◆ physically related to
 - ◆ spatially related to
 - ◆ temporally related to
 - ◆ functionally related to
 - ◆ conceptually related to



- SPECIALIST Lexicon
 - ◆ English dictionary with focus on biomedical terms
 - ◆ Base for the SPECIALIST NLP System
 - ◆ Content: syntactic, morphological und orthographic information

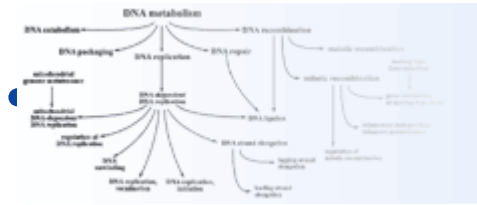
- SPECIALIST NLP Tools
 - ◆ Management of lexical variants and text analysis in the context of biomedicine
 - ◆ Support of integration into information systems



Gene Ontology

- Goal of the GO Project
 - Consistent descriptions of genes/gene products spanning multiple databases
- Three ontologies for description of gene products from different perspectives (overlapping species)
 - ◆ Biological process → 16727 biological process
 - ◆ Cellular component → 2383 cellular component
 - ◆ Molecular function → 8615 molecular function
- Result: uniform queries over the different databases are possible

<http://www.geneontology.org/>



Gene Ontology

- Three-step approach
 - ◆ Development of ontologies
 - ◆ Linkage between ontologies
 - ◆ Development of dedicated tools
- Structure of ontologies: Directed Acyclic Graphs

- Biological process
 - ◆ Series of events, by chaining molecular functions

- Examples
 - ◆ Cellular physiological process
 - ◆ Signal transmission
 - ◆ Pyrimidine metabolism

- A gene product is used in one or multiple biological processes

- Cellular component
 - ◆ should be obvious... 😊
 - ◆ refers to unique, highly organized substances and substances of which cells, and thus living organisms, are composed [Wikipedia]

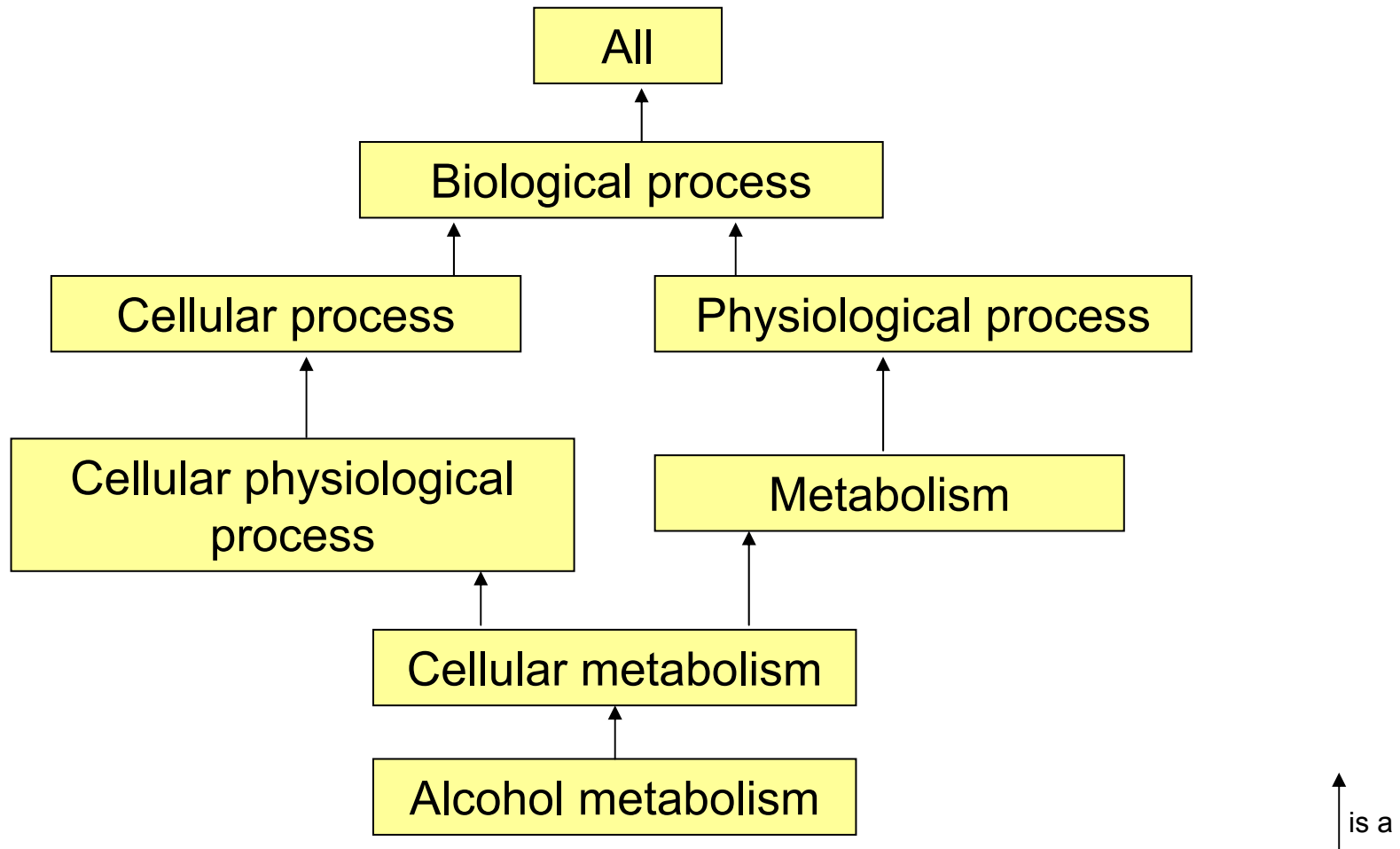
- One gene product is related to one or several cellular components

- Examples:
 - ◆ membranes,
 - ◆ organelles,
 - ◆ proteins,
 - ◆ nucleic acids,

- Molecular function
 - ◆ activities on molecular level, mostly catalytic or binding

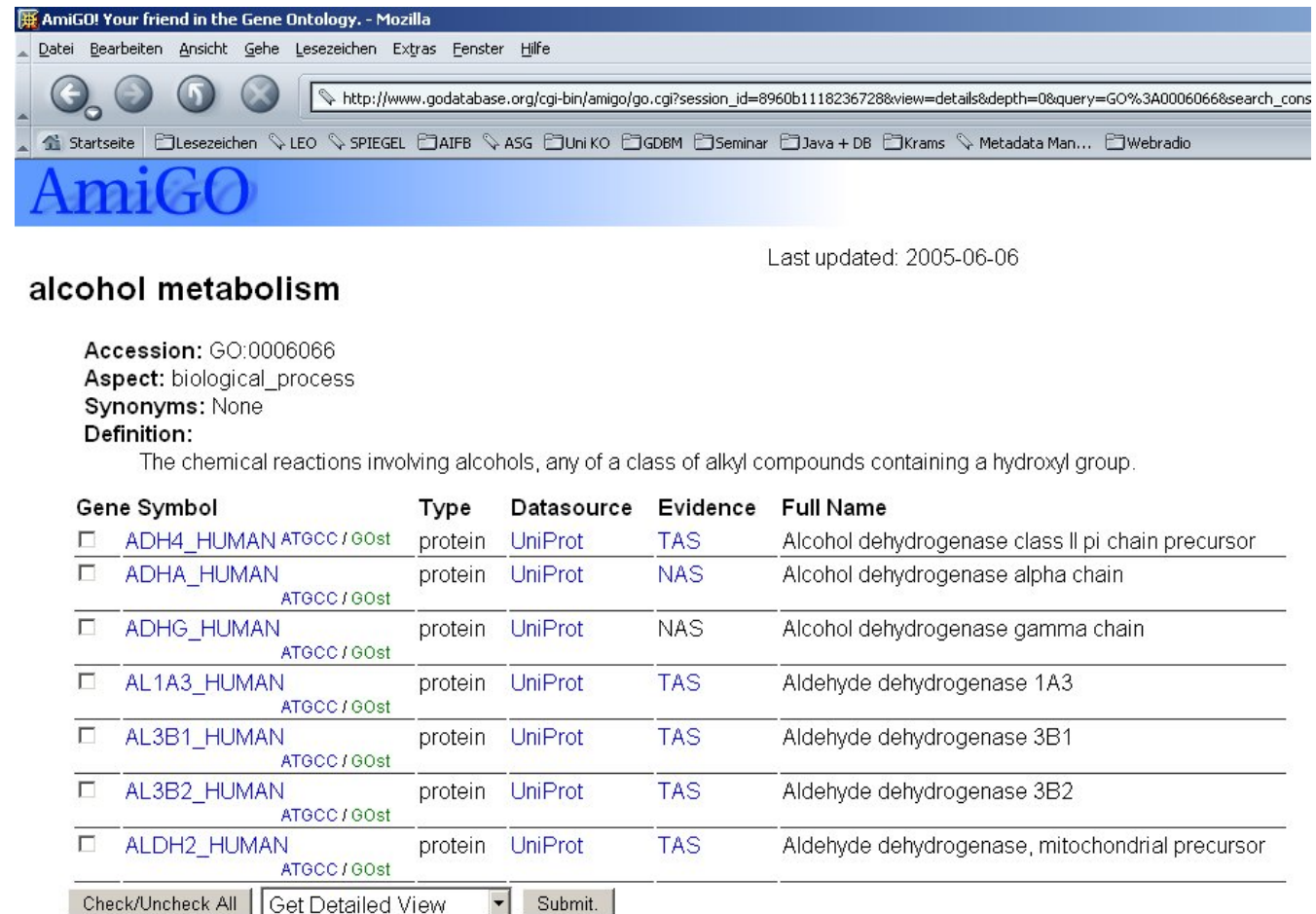
- Examples
 - ◆ Transport
 - ◆ Adenylate cyclase
 - ◆ Phosphate synthase activity

- One gene product has one or multiple molecular functions



Database search, e.g.

- ◆ Gene products with a specific annotation
- ◆ All annotations to one gene product



AmiGO! Your friend in the Gene Ontology. - Mozilla

http://www.godatabase.org/cgi-bin/amigo/go.cgi?session_id=8960b1118236728&view=details&depth=0&query=GO%3A0006066&search_cons

Startseite Lesezeichen LEO SPIEGEL AIFB ASG Uni KO GDBM Seminar Java + DB Krams Metadata Man... Webradio

AmiGO

Last updated: 2005-06-06

alcohol metabolism

Accession: GO:0006066
Aspect: biological_process
Synonyms: None
Definition:
The chemical reactions involving alcohols, any of a class of alkyl compounds containing a hydroxyl group.

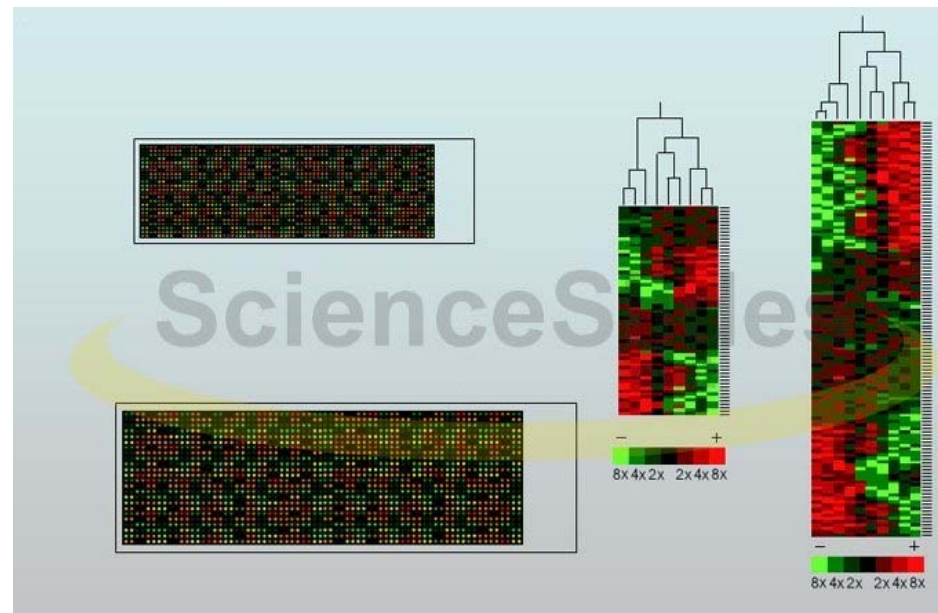
Gene Symbol	Type	Datasource	Evidence	Full Name
<input type="checkbox"/> ADH4_HUMAN <small>ATGCC / GOst</small>	protein	UniProt	TAS	Alcohol dehydrogenase class II pi chain precursor
<input type="checkbox"/> ADHA_HUMAN <small>ATGCC / GOst</small>	protein	UniProt	NAS	Alcohol dehydrogenase alpha chain
<input type="checkbox"/> ADHG_HUMAN <small>ATGCC / GOst</small>	protein	UniProt	NAS	Alcohol dehydrogenase gamma chain
<input type="checkbox"/> AL1A3_HUMAN <small>ATGCC / GOst</small>	protein	UniProt	TAS	Aldehyde dehydrogenase 1A3
<input type="checkbox"/> AL3B1_HUMAN <small>ATGCC / GOst</small>	protein	UniProt	TAS	Aldehyde dehydrogenase 3B1
<input type="checkbox"/> AL3B2_HUMAN <small>ATGCC / GOst</small>	protein	UniProt	TAS	Aldehyde dehydrogenase 3B2
<input type="checkbox"/> ALDH2_HUMAN <small>ATGCC / GOst</small>	protein	UniProt	TAS	Aldehyde dehydrogenase, mitochondrial precursor

Check/Uncheck All Get Detailed View Submit.

GoFigure: Predicting functionality of gene sequences

- GO annotated DBs contain gene sequences
- BLAST algorithm: identification of annotated homologues to unknown gene sequences

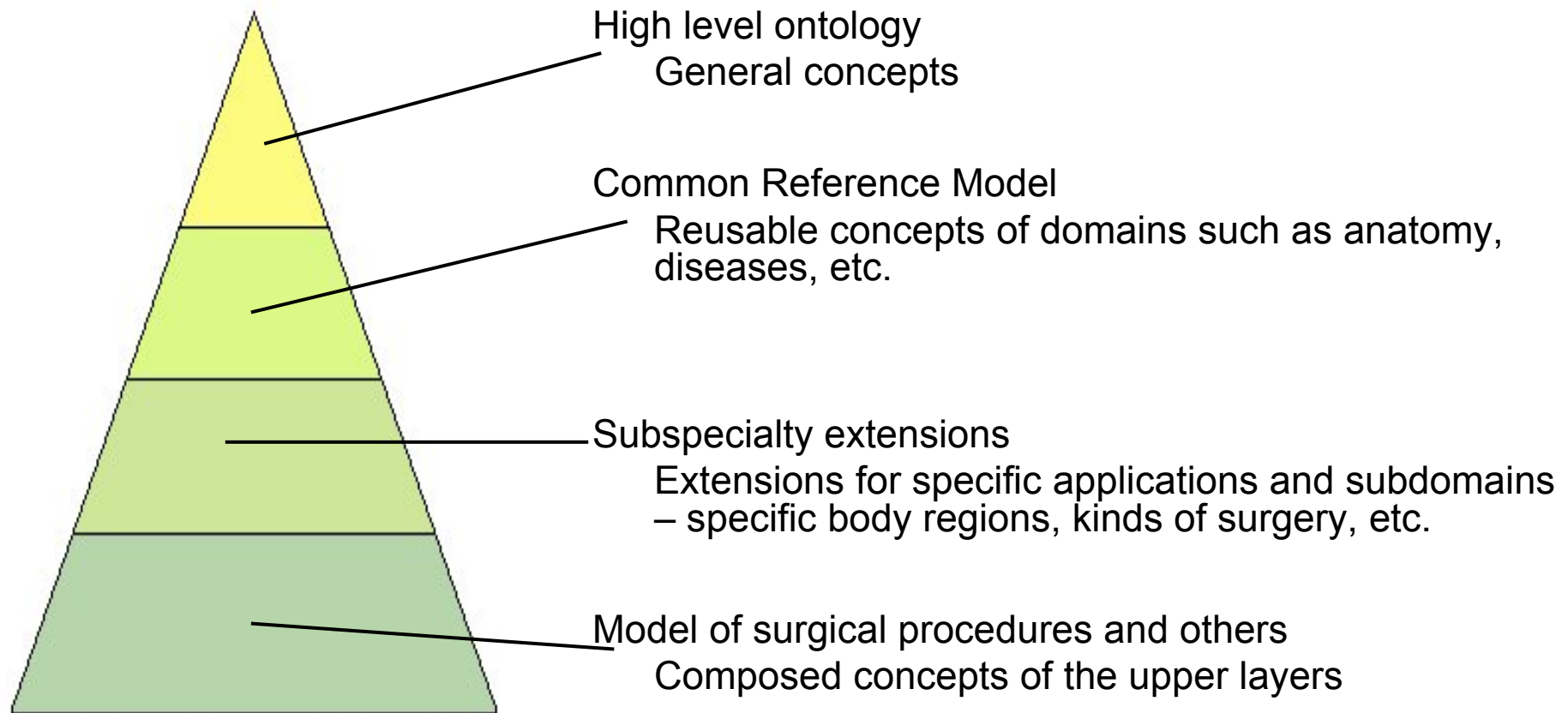
Cancer research: Functional clustering of expressed genes in tumor cells (Microarray analysis)



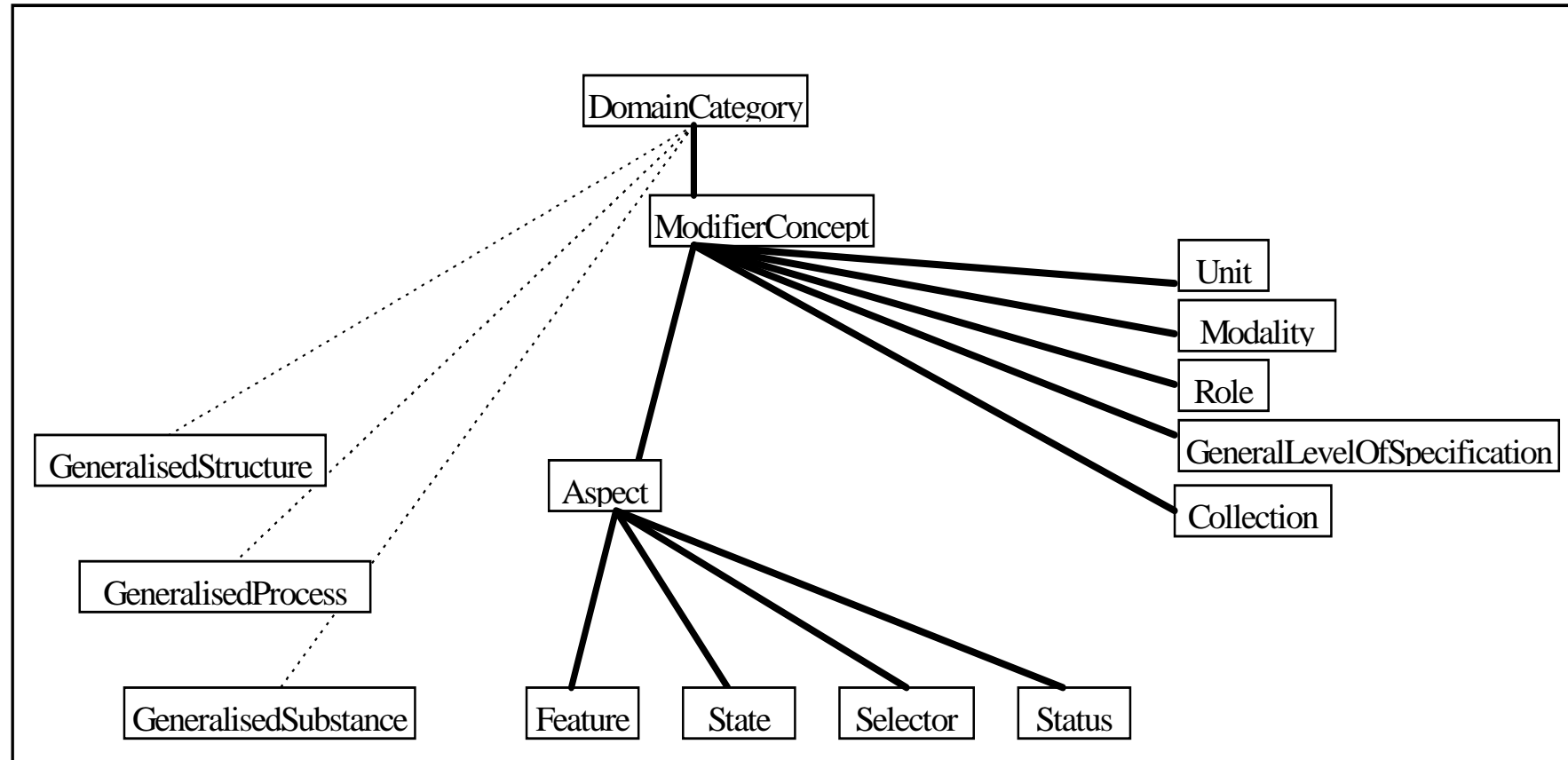
- **Generalized Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine**
- Infrastructure for the integration of terminology in clinical information systems
- Concept system independent of natural language and used coding system
- Publisher: OpenGalen – Non profit organization (U. Manchester, U. Nijmegen)

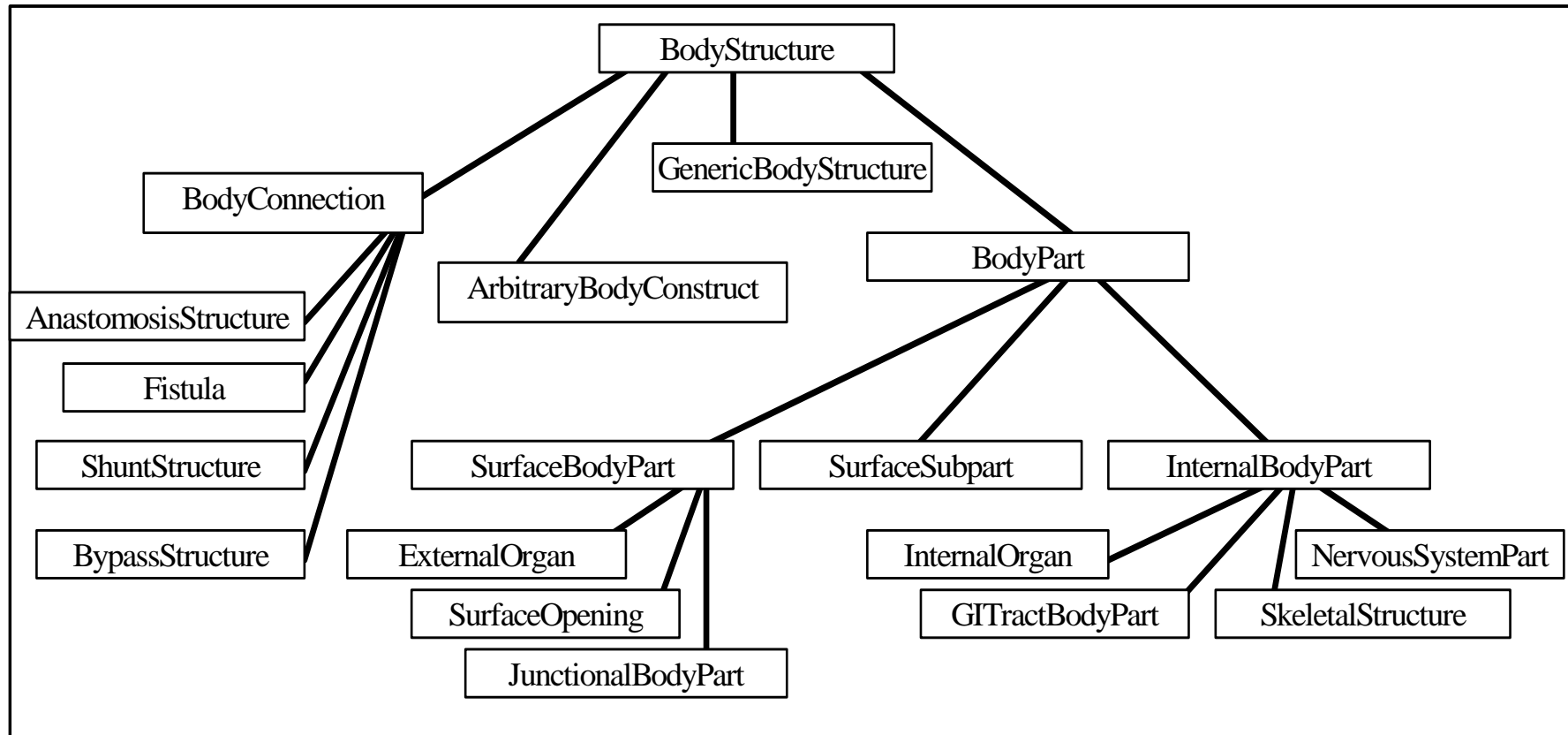
<http://www.opengalen.org/>

- Representation of formal definitions of concepts and their relations
- Move from simple enumeration to composition of concepts
 - Separation of grammar and lexicon
- Realization:
 - ◆ Terminology: GALEN Common Reference Model (CORE)
 - ◆ Language: GALEN Concept Representation Language (GRAIL)
 - ◆ Implementation: GALEN Terminology Server



High level ontology (Upper parts of the hierarchy)





Extension by additional hierarchy axes by attributes, e.g.

- ◆ hasTopology
- ◆ hasSeverity
- ◆ hasShape
- ◆ HasDivision
- ◆ hasSolidRegion
- ◆ hasLeftRightSelector
- ◆ hasStructuralComponent
- ◆ isMadeOf
- ◆ Contains
- ◆ PassesThrough

Example

```
(HeartAtrium which hasLeftRightSelector rightSelection)  
  name RightHeartAtrium.
```

```
(HeartVentricle which hasLeftRightSelector rightSelection)  
  name RightHeartVentricle.
```

```
RightSideOfHeart sensiblyAndNecessarily  
  hasSpecificStructuralComponent [RightHeartVentricle,  
  RightHeartAtrium].
```

- **GALEN Representation And Integration Language**
 - “... formal system, which we use to represent all and only sensible medical concepts [in machine readable format]”
- Formal language for expressing the CORE Model
- Based on Description Logics and Conceptual Graphs
- Additional elements for
 - ◆ Extension of the classification (‘essential criteria’, ‘necessary statements’)
 - ◆ Embedding of definitions
 - ◆ Uniform classification of categories and individuals (no A-Box)
 - ◆ Constraints on negation, disjunction, equality, quantification und referencing

Access to the encapsulated CORE Model via several service interfaces

- ◆ Concept services: classification of concepts into a hierarchy
- ◆ Language services: transformation between concepts and natural language terms
- ◆ Coding services: discovery of codes in classification systems for a concept
- ◆ Indexing services: access to detailed information about a specific concept

- UMLS
<http://umlsinfo.nlm.nih.gov/>
- Gene Ontology
<http://www.geneontology.org/>
<http://udgenome.ags.udel.edu/gofigure/>
- Galen
<http://www.opengalen.org/>