Ontologien in den Lebenswissenschaften

Seminar Ontologie-Management

2. Februar 2009

Universität Leipzig
Abteilung Datenbanken

Bearbeiter: Marcus Stuber
Betreuerin: Anika Groß

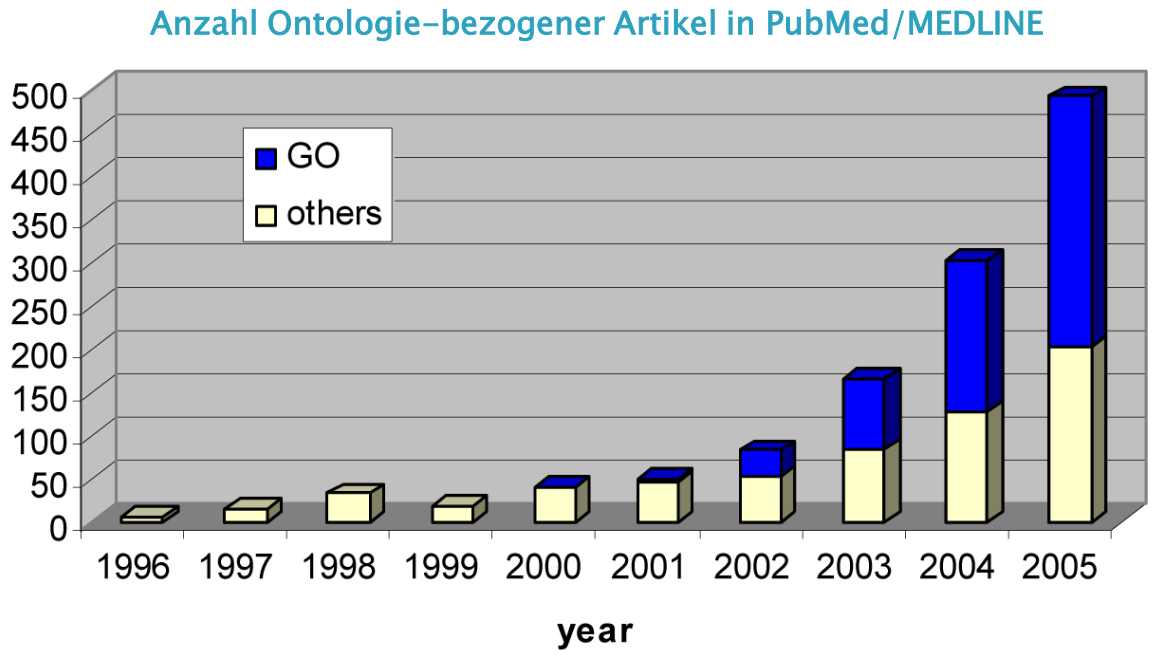
Inhalt

Einleitung

- biologische und medizinische Forschung erzeugen riesige Datenmengen
- benötigt technische Hilfsmittel um diese Informationen durchsuchen und verarbeiten zu können
- Problem: Daten sind heterogen
  - unterschiedliche Terminologien
    Beispiel Glukoseproduktion: *glucose synthesis*, *glucose biosynthesis*, *glucose formation*, *glucose anabolism*, *glucoseneogenesis*
  - unterschiedliche Datenformate und Zugriffsmethoden
  - Wissen in natürlicher Sprache notiert

- benötigt Standard mit definierten Vokabular, der in der Lage ist, auch die Semantik der Informationen formal darzustellen
  ⇒ Ontologien

- Ontologien haben massiv an Bedeutung gewonnen



Ontologien in den Lebenswissenschaften

- Ontologien unterscheiden sich hinsichtlich ihres Aufbaus und ihrer Komplexität

- ⇒ verschiedene Anwendungsbereiche

  - Abfrage heterogener Daten
  - Datenaustausch zwischen Anwendungen
  - Informationsintegration
  - Repräsentation enzyklopädischen Wissens
  - Reasoning
  - Natural Language Processing

Ontologien in den Lebenswissenschaften

Ontologien in der Biologie

- spielen besonders in der Forschung eine wichtige Rolle
  Beispiele: Gene Ontology, Sequence Ontology
- Entwicklung der Open Biomedical Ontologies (OBO)
  die OBO decken viele Gebiete ab, u.a. Anatomie, Biochemie, …
- Entwicklung von Ontologien zur Beschreibung von Experimenten
  Beispiele: Microarray Gene Expression Data (MGED), FuGO

Ontologien in den Lebenswissenschaften

Ontologien in der Medizin

- dienen häufig primär dazu, eine standardisierte Terminologie bereitzustellen
- fehlerfreie Kommunikation ist besonders wichtig in der Medizin
- Beispiele: SNOMED-CT, ICD, UMLS
- es gibt auch Ontologien zur Inhaltserschließung biomedizinischer Literatur
  Beispiel: Medical Subject Headings (MeSH)

Bedeutende Ontologien

Open Biomedical Ontologies

- Entwicklung zahlreicher weiterer Ontologien
  ⇒ Ontologien würden untereinander wieder syntaktisch und semantisch heterogen sein

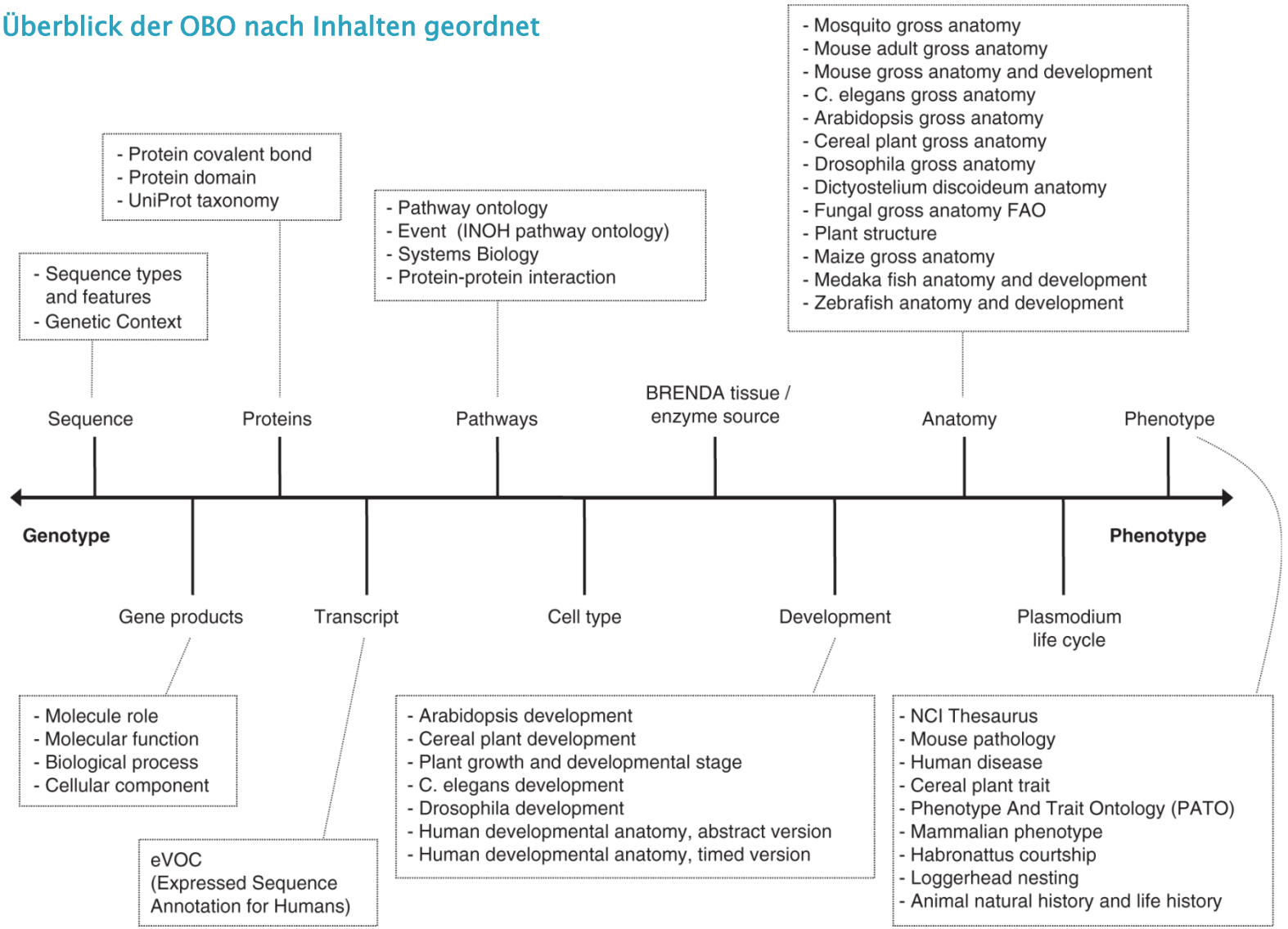
- ⇒ 2001 Gründung des OBO Konsortiums als Dachorganisation für die Entwicklung von Bio-Ontologien

- Einführung einer Reihe von Richtlinien:
  - alle Ontologien müssen offen und frei zugänglich sein
  - alle Ontologien nutzten eine gemeinsame Repräsentation, entweder OWL oder das OBO Format
  - jeder Begriff besitzt neben seinem Namen einen eindeutigen Identifikator
  - alle Ontologien haben einen klar spezifizierten und abgegrenzten Inhalt
  - alle Ontologien sind orthogonal zueinander
  - alle Begriffe besitzen eine textuelle Definition
  - alle Ontologien sind wohldokumentiert

- die OBO beinhalten über 60 Ontologien, die viele verschiedene Bereiche der Biologie abdecken

  - Anatomie
  - Biochemie
  - biologische Funktionen
  - biologische Prozesse
  - Experimente
  - Gesundheit
  - Phänotypen
  - Proteine
  - Sequenzen
  - Taxonomie
  - Umwelt

Überblick der OBO nach Inhalten geordnet



- nahezu alle OBO liegen sowohl in OWL als auch im OBO Format vor
- OBO Flat File Format ist eine Ontologie-Repräsentationssprache, die semantisch OWL-DL ähnelt

OBO-Relation	OWL-Relation
is_a	owl:subClassOf/owl:subPropertyOf
disjoint_from	owl:disjointWith
union_of	owl:unionOf
inverse_of	owl:inverseOf

- OBO Format ist aber einfacher aufgebaut und besitzt einige Erweiterungen

```
[Term]
id: SO:0000087
name: nuclear_gene
def: "A gene from nuclear sequence." [SO:xp]
synonym: "nuclear gene" EXACT []
is_a: SO:0000704 ! implied link automatically realized ! gene
intersection_of: SO:0000704 ! gene
intersection_of: has_origin SO:0000738 ! nuclear_sequence
relationship: has_origin SO:0000738 ! implied link automatically realized ! nuclear_sequence
```

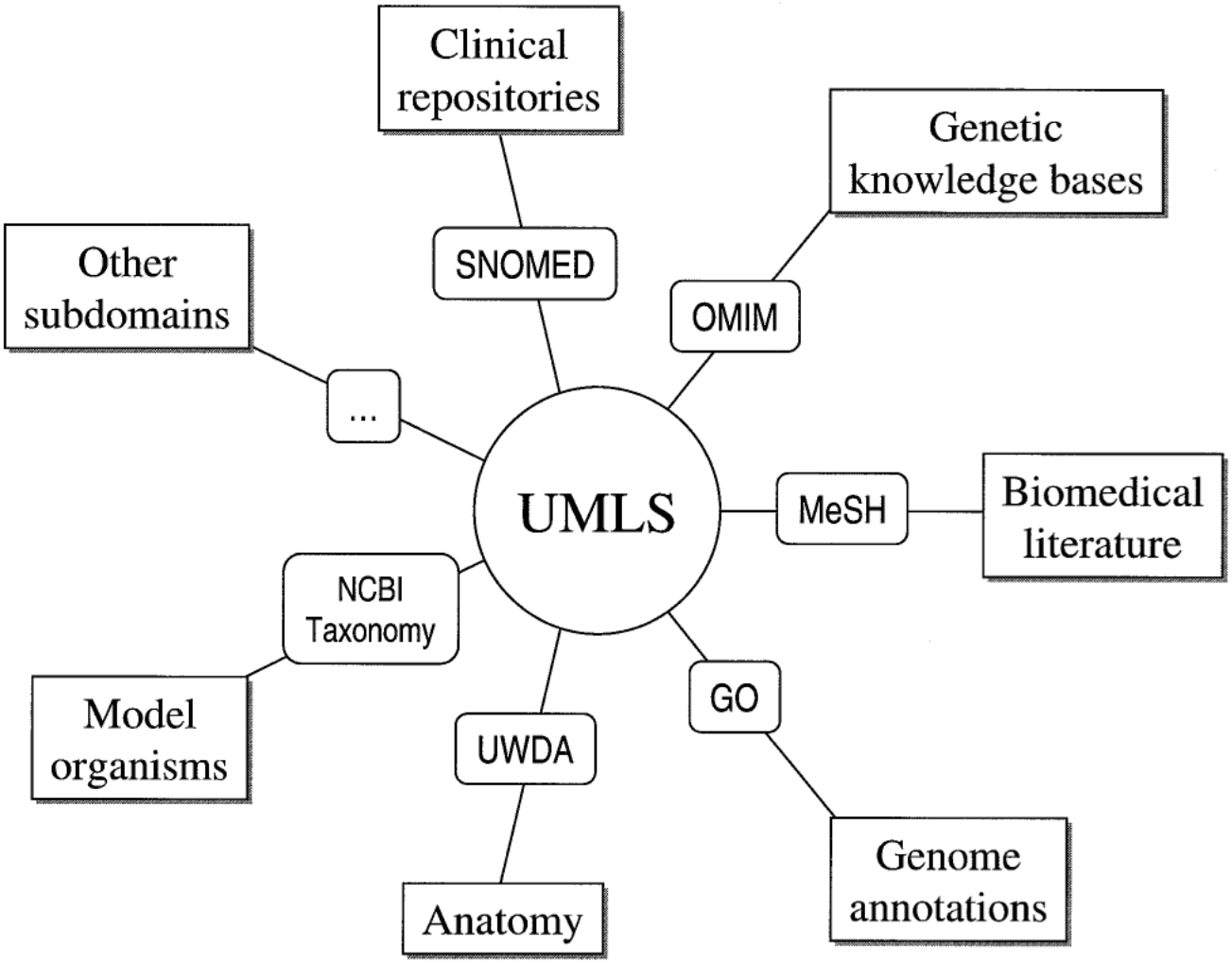
- OBO Format unterstützt Versionierung durch Versions-Tag
- genaue Regelung ist aber den Ontologie-Anbietern überlassen
- alle Ontologien sind in Form eines Graphen strukturiert
- Relationen wurden anfangs noch inkonsistent genutzt
  ⇒ Entwicklung der OBO Relation Ontology

Unified Medical Language System

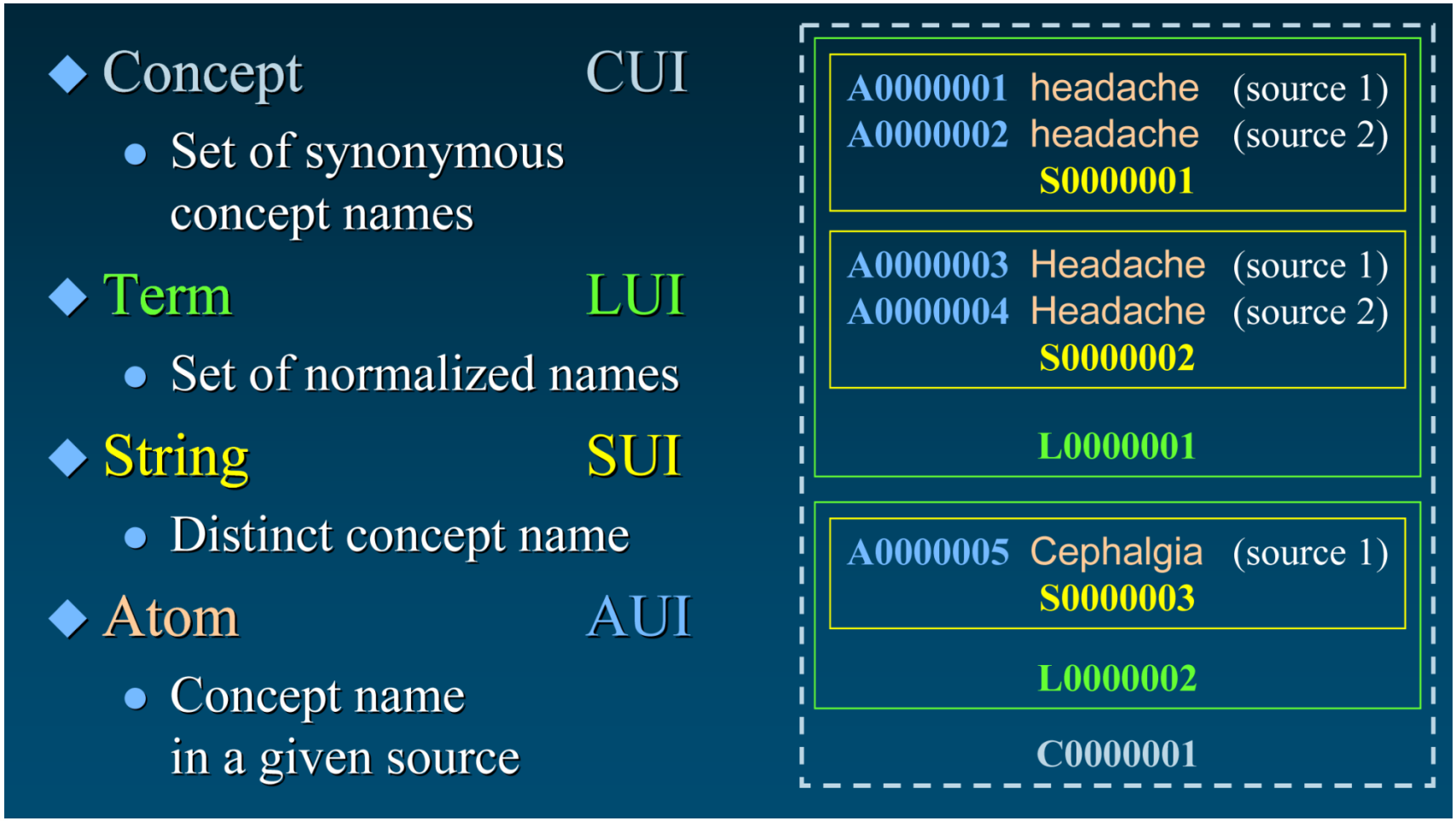
- seit 1986 von der U.S. National Library of Health entwickelt
- versucht
  - die unterschiedlichen Terminologien in einer einzelnen Ontologie zu integrieren
  - Verteilung von Informationen zwischen verschiedenen Datenbanken und Systemen zu ermöglichen
- Fokus liegt auf medizinischen Inhalten
- UMLS besteht aus drei Teilen:
  - dem Metathesaurus
  - dem Semantic Network
  - dem SPECIALIST Lexicon und lexikalischen Werkzeugen

Der Metathesaurus

- fasst alle Konzepte und Begriffe der Quellen zusammen
- enthält Informationen über
  - die diversen biomedizinischen Konzepte
  - ihre verschiedenen Namen
  - und ihre Beziehungen untereinander
- Themenbereich des Metathesaurus wird durch die Inhalte der Quellvokabulare bestimmt


Clinical repositories
Genetic knowledge bases
Other subdomains
SNOMED
OMIM
...
UMLS
MeSH
Biomedical literature
NCBI Taxonomy
Model organisms
UWDA
GO
Anatomy
Genome annotations

- ein Metathesaurus-Konzept erfasst alle bedeutungsgleichen Konzepte der Quellvokabulare



- verschiedenen Identifikatoren dienen zur effizienten Anpassung des Metathesaurus und zur Erkennung von zeitlichen Änderungen der Konzepte

- Metathesaurus-Konzepte sind über Relationen miteinander verbunden
  Beispiele: part of, caused by, location of

- es gibt Konzept-Attribute, Atom-Attribute und Relations-Attribute

- Metathesaurus beinhaltet Metainformationen über die Eigenschaften der aktuellen und Änderungen zur vorherigen Version

- liegt in zwei Formaten vor: ORF und RRF

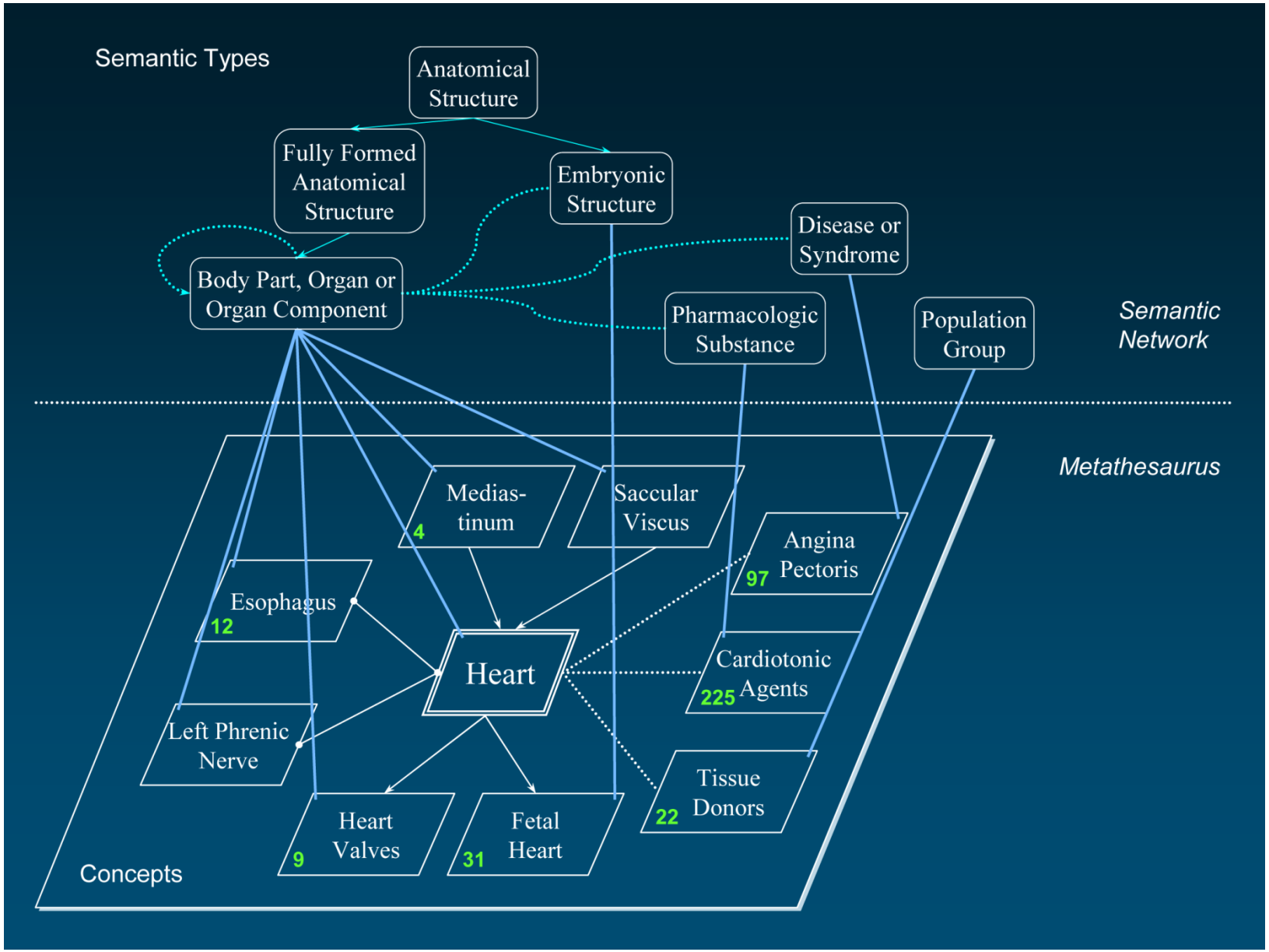
- umfasst über 150 Vokabulare und 1,5 Mio. Konzepte in 17 Sprachen

Das Semantic Network

- dient zur konsistenten Kategorisierung der Konzepte des Metathesaurus
- besteht aus
  - semantischen Typen (Kategorien) z.B. Clinical Drug, Virus, Disease or Syndrome
  - semantischen Relationen z.B. causes, treats
- semantische Typen decken ein breites Spektrum ab
  Beispiele: Organismen, anatomische Strukturen, Chemikalien

Virus causes Disease or Syndrome

- Typen sind über isa hierarchisch strukturiert


Semantic Types
Anatomical Structure
Fully Formed Anatomical Structure
Embryonic Structure
Disease or Syndrome
Body Part, Organ or Organ Component
Pharmacologic Substance
Population Group
Semantic Network
Metathesaurus
Medias-tinum
4
Saccular Viscus
Angina Pectoris
97
Esophagus
12
Heart
Cardiotonic Agents
225
Left Phrenic Nerve
Tissue Donors
22
Heart Valves
9
Fetal Heart
31
Concepts

- jedem Metathesaurus-Konzept wird zumindest ein semantischer Typ zugeordnet
- Semantic Network bildet eine semantische Schicht über dem Metathesaurus
- liegt in zwei Formaten vor: Relational Table Format und Unit Record Format
- 135 semantische Typen und 54 semantische Relationen

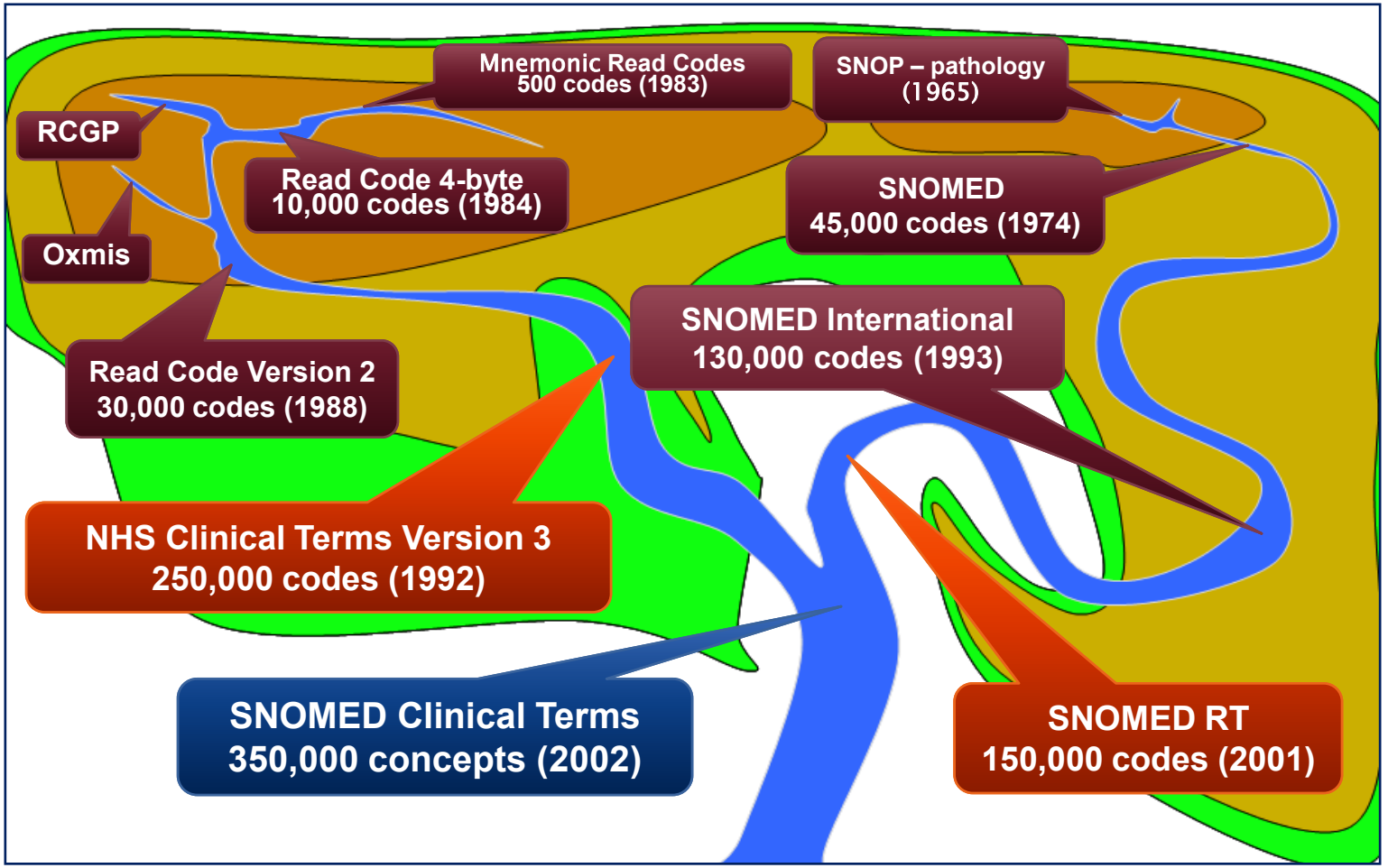
Das SPECIALIST Lexicon

- englischsprachiges Lexikon, das sowohl allgemeine als auch biomedizinische Begriffe umfasst
- entwickelt um notwendige lexikalische Informationen für das SPECIALIST Natural Language Processing System bereitzustellen
- umfasst etwa 300.000 Einträge

- neue Versionen des UMLS erscheinen vierteljährlich
- kostenlos über das Internet zugänglich

SNOMED-CT

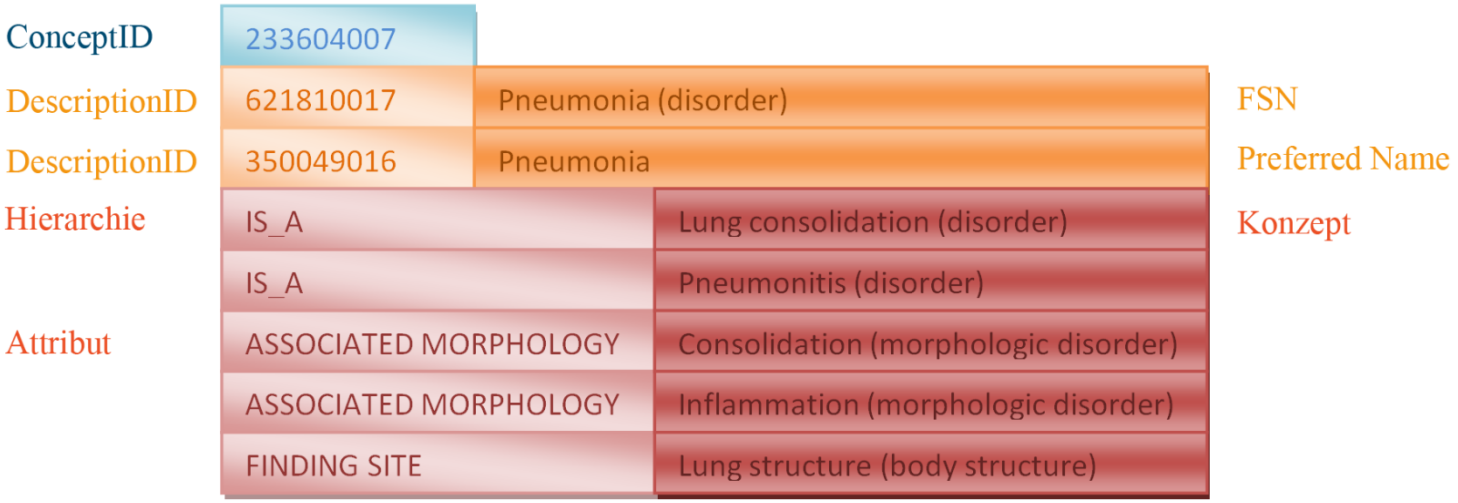
National Health Service College of American Pathologists



- *Systematized Nomenclature of Medicine – Clinical Terms*
- umfassende Terminologie, die Krankheiten, klinische Befunde und Eingriffe erfasst
- ermöglicht konsistente Indexierung, Speicherung, Abfrage, Aggregation und Analyse von Daten
- Vielzahl von Anwendungen
  Beispiele: elektronische Krankenakten, Laborberichte, medizinische Datenbanken
- mittels einer Beschreibungslogik dargestellt

- besteht aus Konzepten, Bezeichnungen und Relationen
- Konzepte besitzen
  - eindeutigen, unveränderlichen, numerischen Identifikator (ConceptID)
  - Reihe von Bezeichnungen
- Bezeichnungen: Synonyme, ein Fully Specified Name (FSN) und ein Preferred Term
- der FSN
  - ist einzigartig
  - dient zu Identifizierung und Erklärung des Konzepts
  - endet mit einem in Klammern geschriebenem Tag
- *Preferred Term* ist geläufige Bezeichnung des Konzepts
- Synonyme sind alle zusätzlichen Begriffe, die das selbe Konzept repräsentieren

- Bezeichnungen besitzen einen Identifikator (DescriptionID)



- Konzepte sind in Hierarchien organisiert

- 19 Top-Level Hierarchien

  - Clinical Finding
  - Procedure
  - Observable Entity
  - Body structure
  - Organism
  - Substance
  - Pharmaceutical/biologic product
  - Specimen
  - Special concept
  - Physical object
  - Physical force
  - Event
  - Environments/geographical locations
  - Social context
  - Situation with explicit context
  - Staging and scales
  - Linkage concept
  - Qualifier value
  - Record artifact

- definierende, modifizierende, historische, ergänzende Relationen

  - definierende: IS_A, Attribute
  - modifizierende: nicht-definierend, schränken Bedeutung weiter ein
  - historische: verbinden inaktive mit aktiven Konzepten
  - ergänzende: zusätzliche nicht-definierende Relationen

- jedes Konzept hat zumindest eine IS_A Beziehung zu einem übergeordneten Konzept und enthält eine Reihe von Attributen

- über 50 verschiedene Attribute

- meisten Attribute auf Konzepte einer Hierarchie beschränkt und dürfen nur auf bestimmte Konzepte verweisen

- SNOMED-CT verfügt über History
- jährlich zwei neue Revisionen der internationalen Version
- kann weitere Terminologien, wie die ICD, referenzieren
- zur Zeit über 311.000 aktive Konzepte
- ist frei über das Internet zugänglich

Medical Subject Headings

- MeSH Thesaurus seit 1960 von der U.S. National Library of Medicine entwickelt
- dient zur Indexierung, Inhaltserschließung biomedizinischer Literatur
- wird u.a. verwendet um Artikel aus 5.200 Fachzeitschriften für MEDLINE zu indexieren
- bibliographische Referenz wird mit einer Reihe von MeSH Begriffen assoziiert, die den Inhalt beschreiben
- Suchanfragen nutzen MeSH Vokabular, um Einträge zu finden

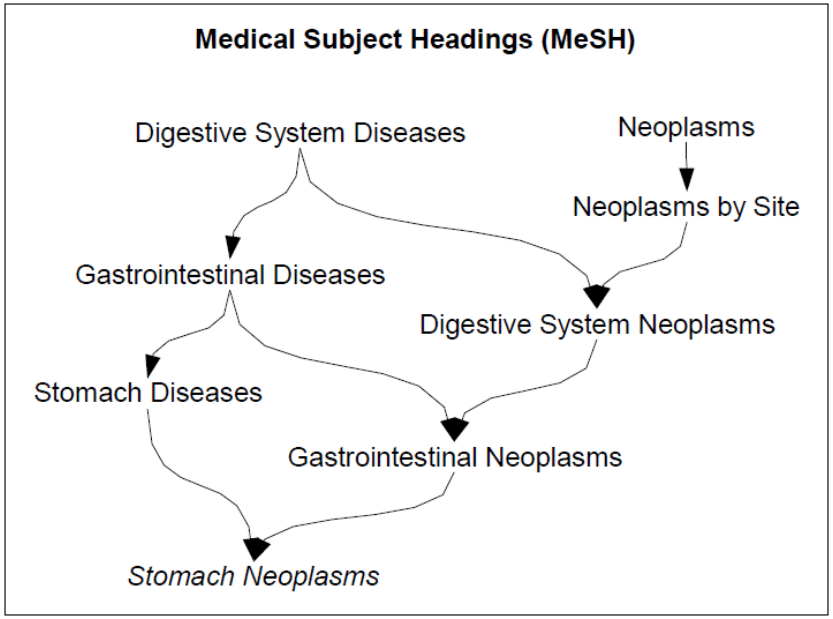
- drei Basistypen von MeSH-Einträgen:
  - Main Headings
  - Subheadings
  - Supplementary Concept Records

- Main Headings
  - dienen zur Beschreibung des Themas eines Dokuments
  - können mehrere Konzepte umfassen, Konzepte beinhalten bedeutungsgleiche Begriffe und textuelle Definition
  - besitzen meist Entry Terms: synonyme oder verwandte Begriffe, die helfen das passende Main Heading zu finden
    Beispiel: Vitamin C ist Entry Term von Main Heading Ascorbic Acid
  - können Kreuzverweise beinhalten
  - alphabetisch und hierarchisch strukturiert

- 16 umfassende Main Headings an der Spitze der Hierarchie

  - Anatomy
  - Organisms
  - Diseases
  - Chemicals and drugs
  - Analytical, Diagnostic and Therapeutic Techniques and Equipment
  - Psychiatry and Psychology
  - Phenomena and Processes
  - Disciplines and Occupations
  - Anthropology, Education, Sociology and Social Phenomena
  - Technology, Industry, Agriculture
  - Humanities
  - Information Science
  - Named Groups
  - Health Care
  - Publication Characteristics
  - Geographicals

- Position eines Main Headings in Hierarchie durch MeSH Tree Number beschrieben, kann mehrere MeSH Tree Numbers besitzen



Digestive System Neoplasms: C06.301 und C04.588.274

- über 25.000 Main Headings mit über 160.000 Entry Terms

- 83 Subheadings: ermöglichen es, Dokumente zu gruppieren, die gewisse Aspekte eines Themas gemeinsam haben
  Beispiel: Liver/drug effects

- über 180.000 Supplementary Concept Records in separatem Thesaurus, dienen zur Indexierung von Substanzen wie Chemikalien oder Medikamenten für MEDLINE

- MeSH existiert in XML und ASCII Format und kann frei über das Internet genutzt oder heruntergeladen werden

- neue Versionen erscheinen jährlich

- in viele Sprachen übersetzt

Zusammenfassung

Name	Anwendung	Umfang	entwickelt	Versions-rhythmus
GO	Annotation von Genprodukten	26.627 Begriffe	1998 Zusammenschluss dreier Datenbanken für Modellorganismen	kontinuierlich
UMLS	Integration biomedizinischer Terminologien	1,5 Millionen Konzepte über 150 Vokabularen, 135 semantische Typen, 54 semantische Relationen, 300.000 Lexikoneinträge	1986 U.S. National Library of Health	vierteljährlich
ICD	Klassifizierung von Krankheiten	12.161 Krankheitsklassen	1893 Jacques Bertillon	unregelmäßig, 10 Jahre und mehr
SNOMED-CT	Klinische Terminologie (Krankenakten)	311.000 Konzepte	2002 britischer National Health Service und College of American Pathologists	2 Revisionen pro Jahr
MeSH	Indexierung biomedizinischer Literatur	25.186 Main Headings, 160.000 Entry Terms, 83 Subheadings, 180.000 SCRs	1960 U.S. National Library of Medicine	jährlich

to be continued…

[1] Daniel Schober: *Ontologien in den Biowissenschaften*.
http://www.bioinf.mdc-berlin.de/~schober/bio-ontologien.htm
[2] Daniel L. Rubin, Nigam H. Shah, Natalya F. Noy: *Biomedical ontologies: a functional perspective*
[3] Olivier Bodenreider, Robert Stevens: Bio-*ontologies: current trends and future directions*
[4] http://www.obofoundry.org/
[5] http://www.ihtsdo.org/snomed-ct/
[6] Olivier Bodenreider: *The Unified Medical Language System (UMLS): integrating biomedical terminology*
[7] http://www.nlm.nih.gov/mesh/2009/introduction/intro_preface.html
[8] Rachel Kleinsorge, Jan Willis: *Unified Medical Language System Basics*
http://www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf
[9] Barry Smith, Michael Ashburner, Cornelius Rosse et al.:
The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration
[10] http://obofoundry.org/wiki/index.php/OBO_Foundry_Principles
[11] http://www.geneontology.org/GO.format.obo-1_2.shtml
[12] http://oboedit.org/docs/html/An_Introduction_to_OBO_Ontologies.htm
[13] Barry Smith, Jacob Köhler, Anand Kumar:
On the application of formal principles to life science data: a case study in the Gene Ontology
[14] http://www.obofoundry.org/cgi-bin/detail.cgi?id=relationship
[15] http://www.nlm.nih.gov/research/umls/meta2.html
[16] Olivier Bodenreider: *UMLS and Semantic Web*
http://mor.nlm.nih.gov/pubs/pres/060321-NIAID.pdf
[17] http://www.nlm.nih.gov/research/umls/meta3.html
[18] http://www.nlm.nih.gov/research/umls/meta4.html
[19] http://www.nlm.nih.gov/pubs/factsheets/umls.html
[20] http://nis-web.fhs.usyd.edu.au/ncch_new/downloads/coding_matters/vol8no2.pdf
[21] SNOMED Clinical Terms and SNOMED Terminology Solutions: A brief overview
http://www.himss.org/content/files/snomed_101overview.pdf
[22] SNOMED Clinical Terms User Guide
http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Technical_Docs/SNOMED_CT_User_Guide_20080731.pdf
[23] J.J. Cimino, X. Zhu: *The Practical Impact of Ontologies on Biomedical Informatics*
[24] http://www.ihtsdo.org/snomed-ct/release-of-snomed-ct/
[25] http://www.nlm.nih.gov/mesh/2009/introduction/intro_preface.html
[26] http://www.nlm.nih.gov/mesh/intro_record_types.html
[27] http://www.nlm.nih.gov/mesh/meshrels.html
[28] http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[29] http://www.nlm.nih.gov/cgi/mesh/2009/MB_cgi
[30] http://en.wikipedia.org/wiki/Medical_Subject_Headings
[31] http://www.nlm.nih.gov/mesh/filelist.html