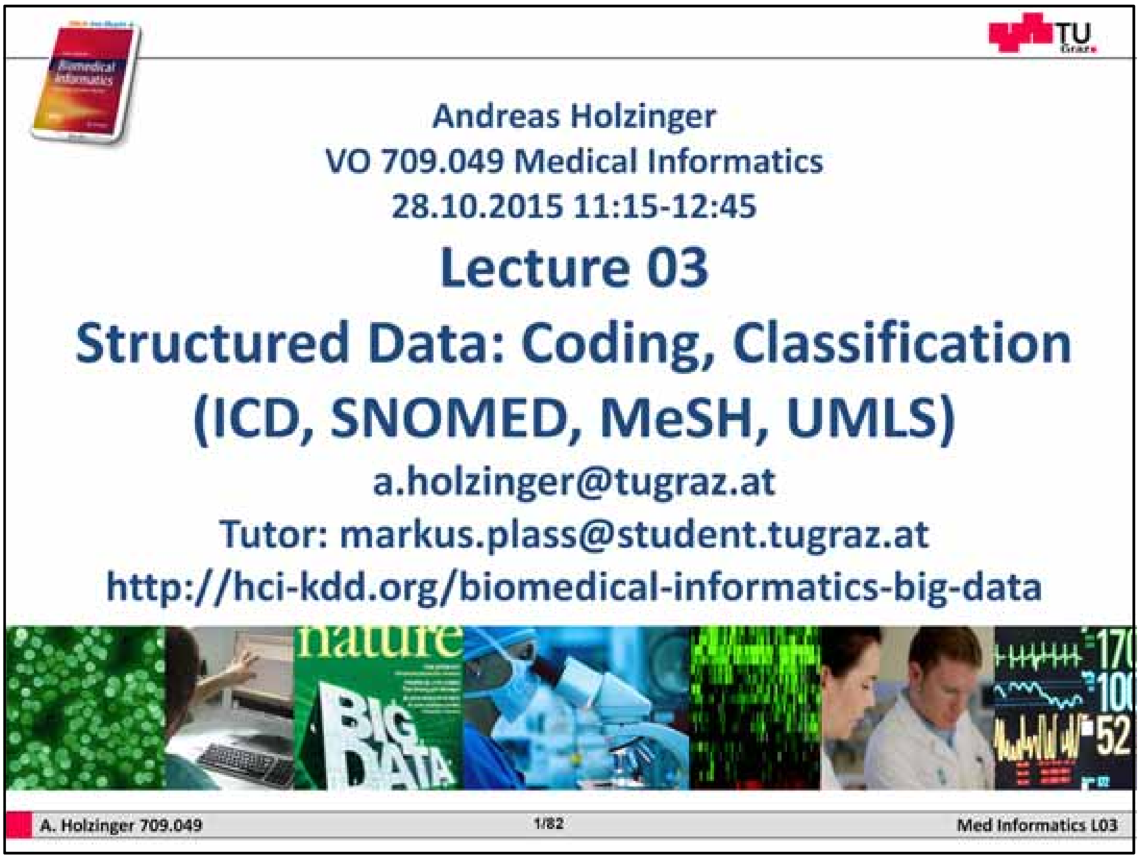


Status as of Mo, 26.10.2015, 10:00

Dear Students, welcome to the 3rd lecture of our course. Please remember from the last lecture: data sources, data structures, standardization versus structurization, the differences notions between data, information and knowledge and close with an overview about information entropy.

Please always be aware of the definition of biomedical informatics (Medizinische Informatik):
Biomedical Infromatics is the inter-disciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health (and well-being).

Schedule

- 1. Intro: Computer Science meets Life Sciences, challenges, future directions
- 2. Back to the future: Fundamentals of Data, Information and Knowledge
- **3. Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)**
- 4. Biomedical Databases: Acquisition, Storage, Information Retrieval and Use
- 5. Semi structured and weakly structured data (structural homologies)
- 6. Multimedia Data Mining and Knowledge Discovery
- 7. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
- 8. Biomedical Decision Making: Reasoning and Decision Support
- 9. Intelligent Information Visualization and Visual Analytics
- 10. Biomedical Information Systems and Medical Knowledge Management
- 11. Biomedical Data: Privacy, Safety and Security
- 12. Methodology for Info Systems: System Design, Usability & Evaluation

A. Holzinger 709.049 2/82 Med Informatics L03

Keywords of the 3th Lecture

- Biomedical Ontologies
- Classification of Diseases
- International Classification of Diseases (ICD)
- Medical Subject Headings (MeSH)
- Modeling biomedical knowledge
- Ontology Languages (OL)
- Resource Description Framework (RDF)
- Standardized Medical Data
- Systematized Nomenclature of Medicine (SNOMED)
- Unified Medical Language System (UMLS)
- Work domain model (WDM)

Learning Goals: At the end of this 3rd lecture you …

- … have acquired background knowledge on some issues in standardization and structurization of data;
- … have a general understanding of modeling knowledge in medicine and biomedical informatics;
- … got some basic knowledge on medical Ontologies and are aware of the limits, restrictions and shortcomings of them;
- … know the basic ideas and the history of the International Classification of Diseases (ICD);
- … have a view on the Standardized Nomenclature of Medicine Clinical Terms (SNOMED CT);
- … have some basic knowledge on Medical Subject Headings (MeSH);
- … understand the fundamentals and principles of the Unified Language System (UMLS);

Advance Organizer (1/2)

- **Abstraction** = process of mapping (biological) processes onto a series of concepts (expressed in mathematical terms);
- **Biological system** = a collection of objects ranging in size from molecules to populations of organisms, which interact in ways that display a collective function or role (= collective behaviour);
- **Coding** = any process of transforming descriptions of medical diagnoses and procedures into standardized code numbers, i.e. to track health conditions and for reimbursement; e.g. based on Diagnosis Related Groups (DRG)
- **Data model** = definition of entities, attributes and their relationships within complex sets of data;
- **DSM** = Diagnostic and Statistical Manual for Mental Disorders
- **Extensible Markup Language (XML)** = set of rules for encoding documents in machine-readable form.
- **GALEN** = Generalized Architecture for Languages, Encyclopedias and Nomenclatures in Medicine is a project aiming at the development of a reference model for medical concepts
- **ICD** = International Classification of Diseases, the archetypical coding system for patient record abstraction (est. 1900)
- **Medical Classification** = provides the terminologies of the medical domain (or at least parts of it), there are 100+ various classifications in use;
- **MeSH** = Medical Subject Headings is a classification to index the world medical literature and forms the basis for UMLS

Advance Organizer (2/2)

- **Metadata** = data that describes the data;
- **Model** = a simplified representation of a process or object, which describes its behaviour under specified conditions (e.g. conceptual model);
- **Nosography** = science of description of diseases;
- **Nosology** = science of classification of diseases;
- **Ontology** = structured description of a domain and formalizes the terminology (concepts-relations, e.g. IS-A relationship provides a taxonomic skeleton), e.g. gene ontology;
- **Ontology engineering** = subfield of knowledge engineering, which studies the methods and methodologies for building ontologies;
- **SNOMED** = Standardized Nomenclature of Medicine, est. 1975, multitaxial system with 11 axes;
- **SNOP** = Systematic Nomenclature of Pathology (on four axes: topography, morphology, etiology, function), basis for SNOMED;
- **System features** = static/dynamic; mechanistic/phenomenological; discrete/continous; deterministic/stochastic; single-scale/multi-scale
- **Terminology** = includes well-defined terms and usage;
- **UMLS** = Unified Medical Language System is a long-term project to develop resources for the support of intelligent information retrieval;

A. Holzinger 709.049 6/82 Med Informatics L03

Glossary

- ACR = American College of Radiologists
- API = Application Programming Interface
- DAML = DARPA Agent Markup Language
- DICOM = Digital Imaging and Communications in Medicine
- DL = Description Logic
- ECG = Electrocardiogram
- EHR = Electronic Health Record
- FMA = Foundational Model of Anatomy
- FOL = First-order logic
- GO = Gene Ontology
- ICD = International Classification of Diseases
- IOM = Institute of Medicine
- KIF = Knowledge Interchange Format, a FOL-based language for knowledge interchange.
- LOINC = Logical Observation Identifiers Names and Codes
- MeSH = Medical Subject Headings
- MRI = Magnetic Resonance Imaging
- NCI = National Cancer Institute (US)
- NEMA = National Electrical Manufacturer Association
- OIL = Ontology Inference Layer (description logic)
- OWL = Ontology Web Language
- RDF = Resource Description Framework
- RDF Schema = A vocabulary of properties and classes added to RDF
- SCP = Standard Communications Protocol
- SNOMED CT = Systematized Nomenclature of Medicine – Clinical Terms
- SOP = Standard Operating Procedure
- UMLS = Unified Medical Language System

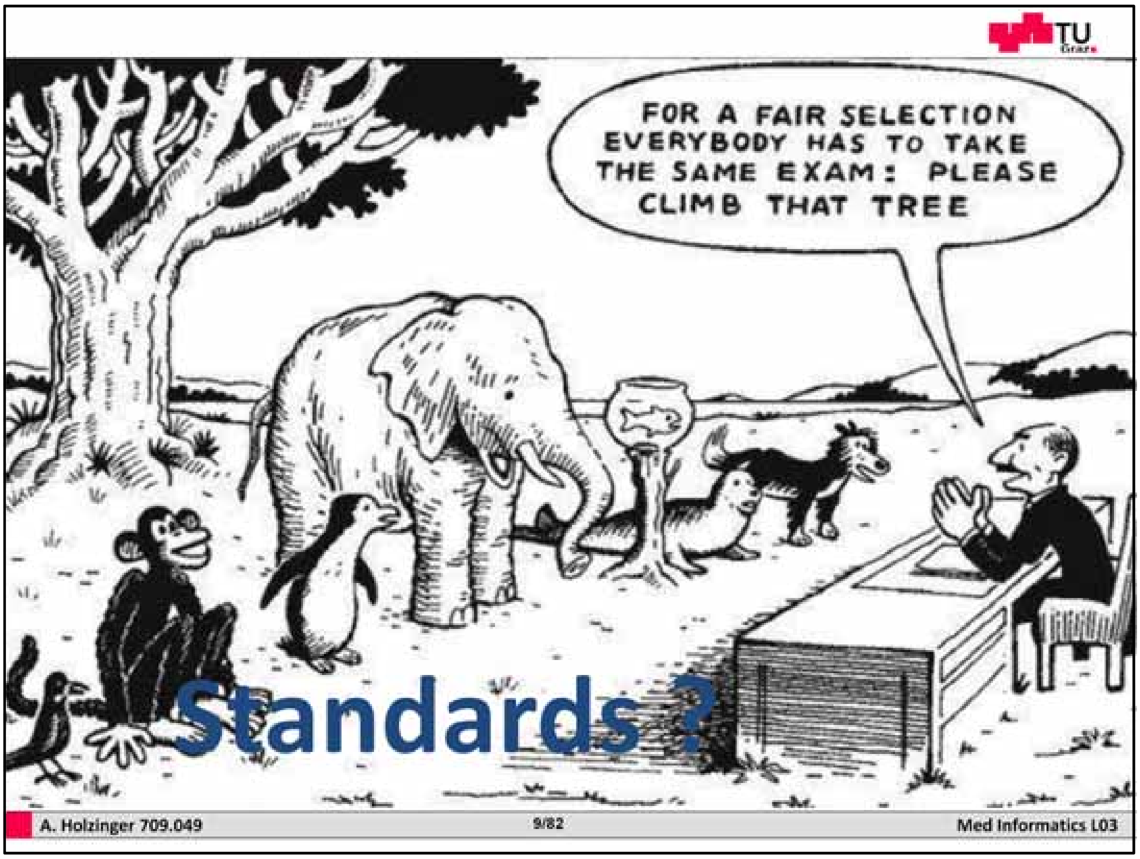
Key Problems

- To find a trade-off between standardization and **personalization** [1];
- The large amounts of **non-standardized data** and **unstructured information** ("free text") [2];
- **Low integration** of standardized terminologies in the daily clinical practice (Who is using e.g. SNOMED, MeSH, UMLS in daily routine?);
- **Low acceptance** of classification codes amongst practitioners;

1. Holmes, C., Mcdonald, F., Jones, M., Ozdemir, V., Graham, J. E. 2010. Standardization and Omics Science: Technical and Social Dimensions Are Inseparable and Demand Symmetrical Study. Omics-Journal of Integr. Biology, 14, (3), 327-332.
2. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C. & Verspoor, K. 2014. Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges. In: LNCS 8401. Berlin Heidelberg: Springer pp. 271-300.

Holmes, C., Mcdonald, F., Jones, M., Ozdemir, V., Graham, J. E. 2010. Standardization and Omics Science: Technical and Social Dimensions Are Inseparable and Demand Symmetrical Study. Omics-Journal of Integrative Biology, 14, (3), 327-332.

On how to deal with unstructured information:
Mack, R., Mukherjea, S., Soffer, A., Uramoto, N., Brown, E., Coden, A., Cooper, J., Inokuchi, A., Iyer, B., Mass, Y., Matsuzawa, H. & Subramaniam, L. V. 2004. Text analytics for life science using the unstructured information management architecture. IBM Systems Journal, 43, (3), 490-515.

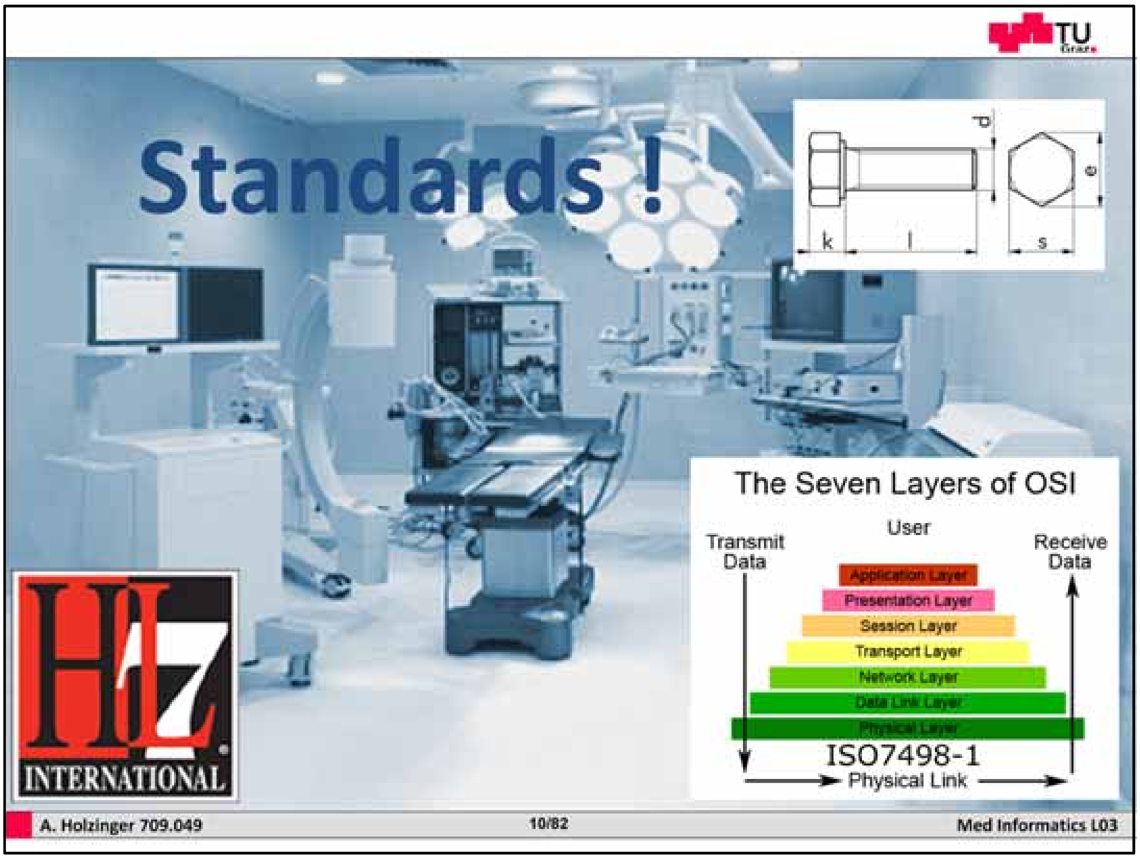


http://jansimson.files.wordpress.com/2012/09/education.png

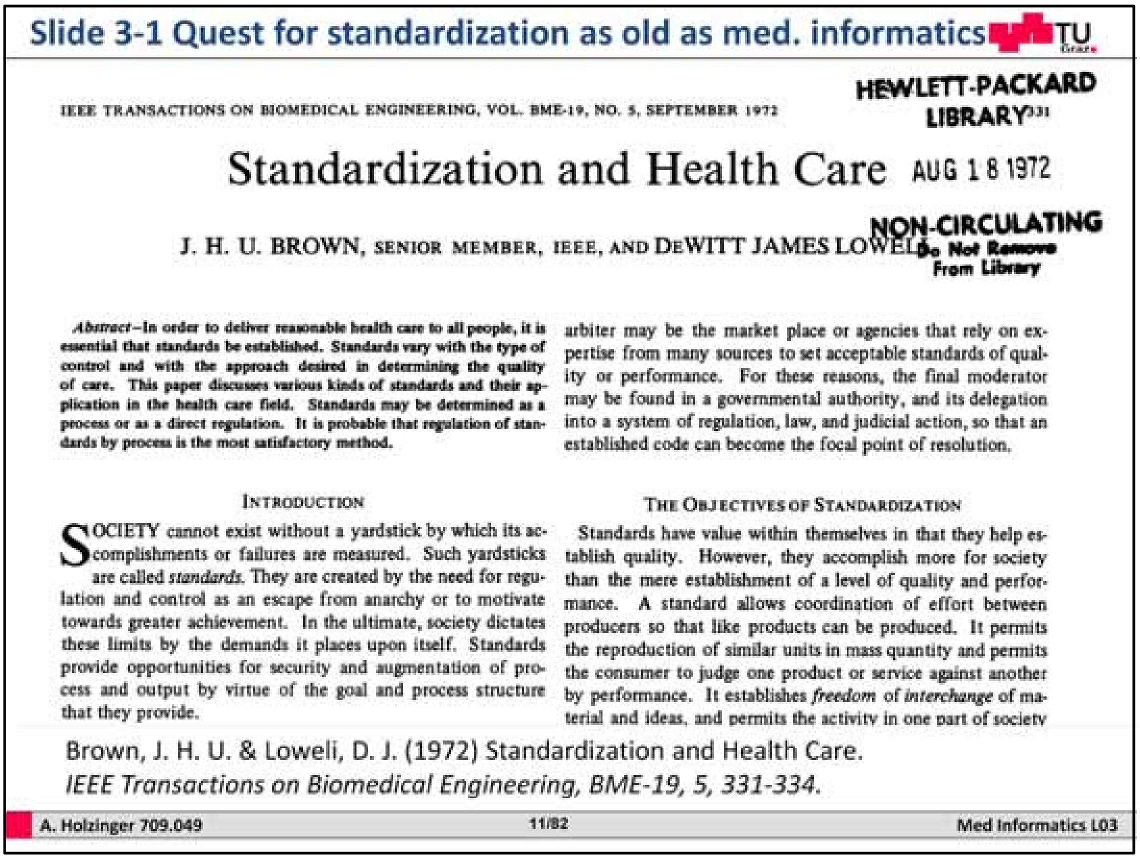
A technical standard is an established norm specified in a formal document and valid on the basis of convention.

The ISO metric screw threads are the world-wide most commonly used type of general-purpose screws. They were one of the first international standards agreed when the International Organization for Standardization was set up in 1947.
The "M" designation for metric screws indicates the nominal outer diameter of the screw, in millimeters (e.g. an M5 screw has a nominal outer diameter of 5 millimeters).
The screw thread was invented around 400 BC by Archytas of Tarentum (founder of mechanics and a contemporary of Plato).

Legido-Quigley, H., Mckee, M., Walshe, K., Sunol, R., Nolte, E. & Klazinga, N. 2008. How can quality of health care be safeguarded across the European Union? British Medical Journal, 336, (7650), 920-923.



A grand challenge in medicine and healthcare is complexity. Standardization is a systematic approach to create order, making selections, and formulating rules and practices. Consequently, it is indispensable for creating context (using the same terminologies, vocabularies etc.), exchange data, provide standard operating procedures (SOP's) and enable interoperability of devices. We define:
Standard is a recognized norm that establishes criteria, methods, processes, practices, etc. which lead to interoperability, compatibility, and repeatability. Note: The existence of a published and recognized standard does not necessarily imply that it is useful or correct. For practical purposes only two generic types of standards exist: standards of quality and standards of production (aka standards of quantity).
Standards of quality are measured by the attributes or properties of a product, material, process etc., which defines the goals of a desired performance. Standards of production refer to the execution of a repeated process not necessarily characterized by product quality as much as by end-product reproducibility. Both standards have high value for a health care system (Brown & Loweli, 1972).
Standardization is the process of developing and implementing standards.
An example is the Evidence Based Medicine (EBM) approach, using techniques from science, engineering and statistics, including systematic review of medical literature, meta-analysis, risk-benefit analyses, and randomized controlled trials (RCTs). This quality approach aims for the ideal that healthcare professionals should make "conscientious, explicit, and judicious use of current best evidence" in their everyday practice.

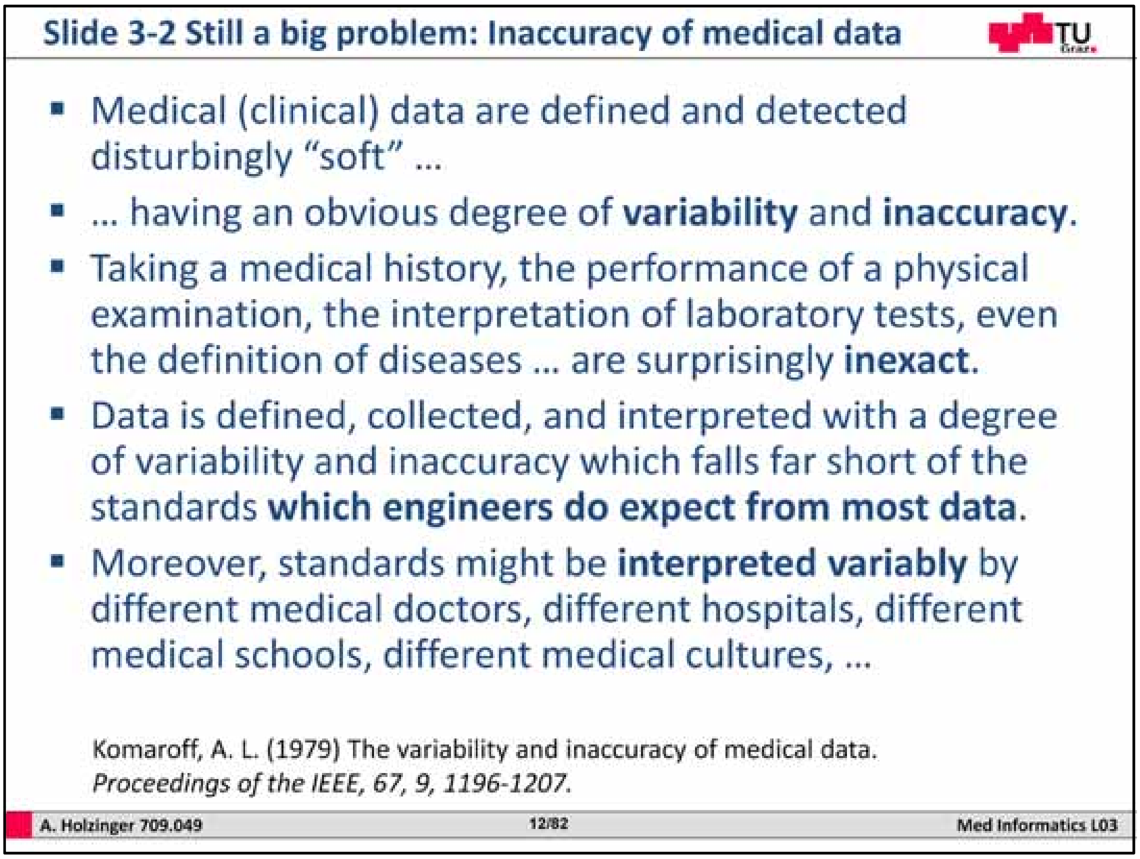


Brown & Loweli (1972) describe the need for standards in order to deliver reasonable health care to all people – at a time when medical informatics was in its infancy and electronic patient records were still science fiction.

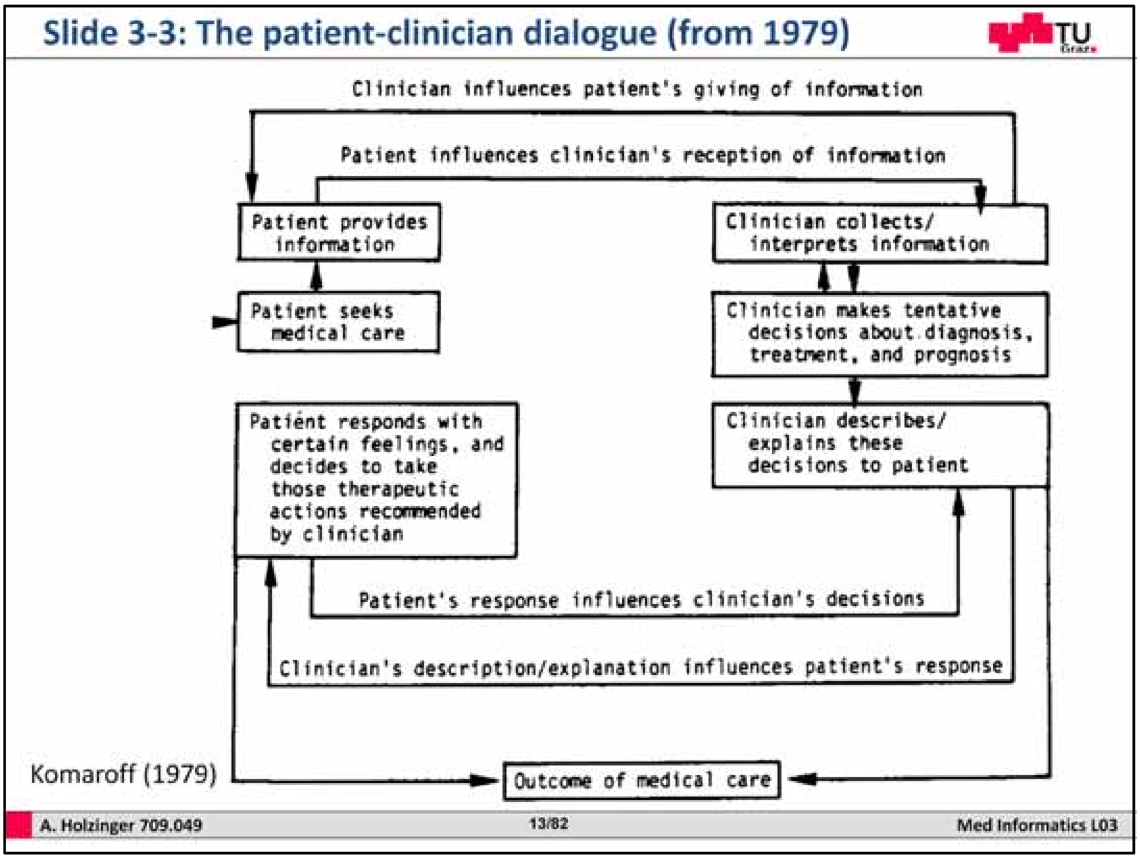
EFQM

The European Foundation for Quality Management (EFQM) provides a framework for self assessment that is used by facilities applying for the European Quality Award and corresponding national awards. EFQM was founded in 1988 by the presidents of 14 major European companies, with the endorsement of the European Commission. It seeks to stimulate and help organisations participate in improvement activities, leading to excellence in customer and employee satisfaction, and thus an impact on society and business performance.It follows the Donabaedian structure-process-outcome principle and emphasises organisational development through self assessment. Two elements, “positioning and improving” and “self-assessment,” are especially relevant to healthcare organisations.

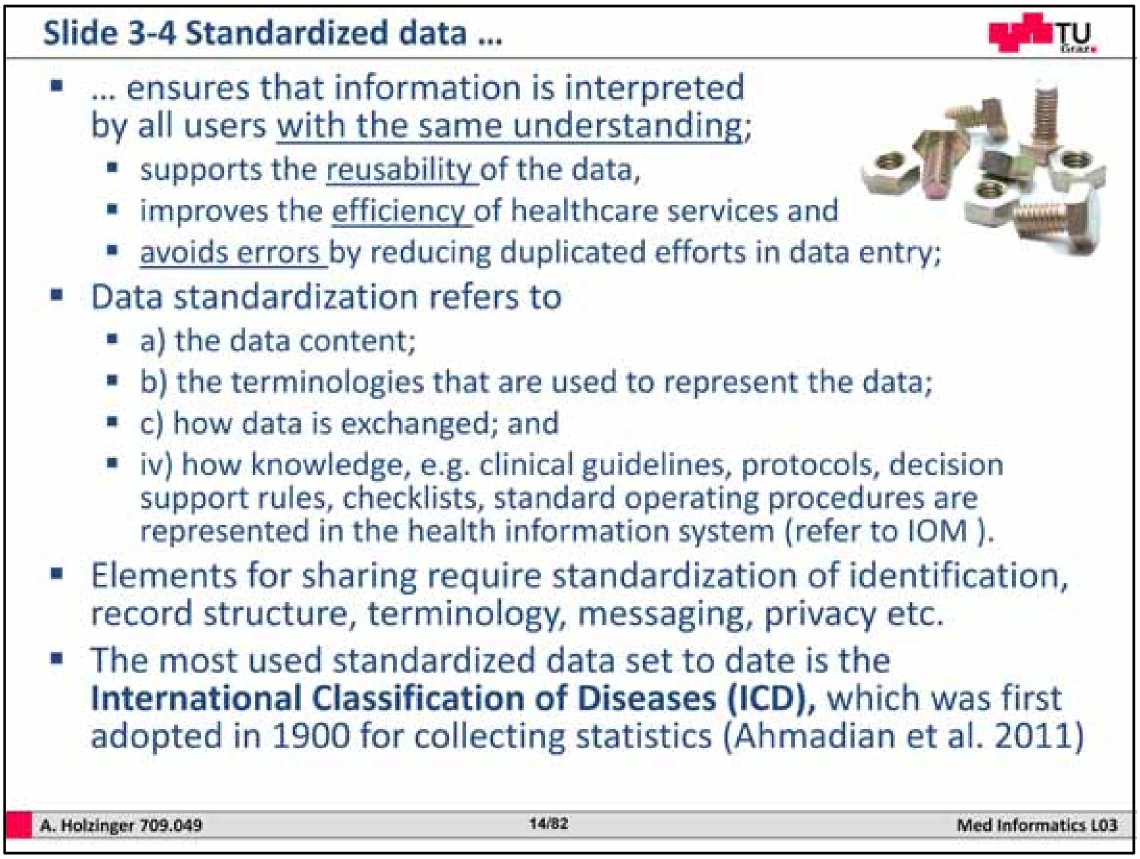
Klazinga N. Re-engineering trust: the adoption and adaption of four models for external quality assurance of health care services in western European health care systems. Int J Qual Health Care 2000;12:183-9.



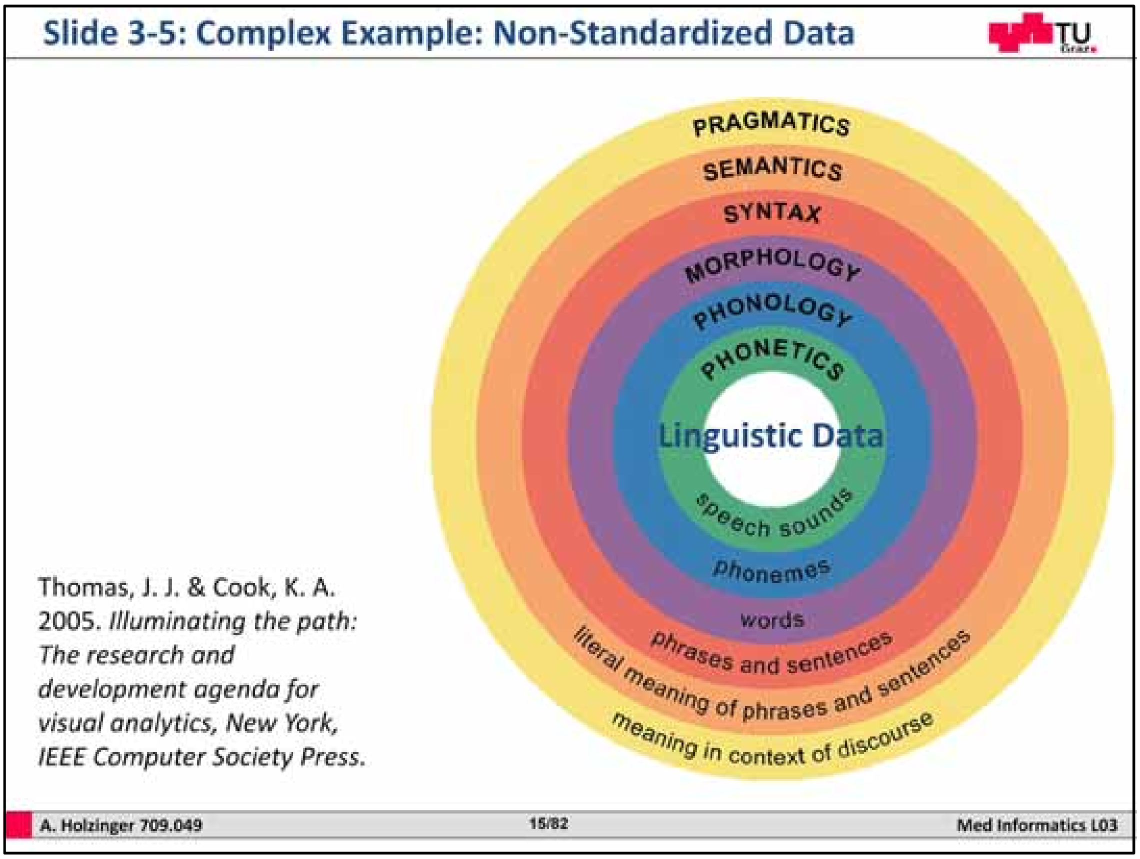
Komaroff (1979) describes clinical data as being disturbingly "soft", having an obvious degree of variability and inaccuracy. Taking a medical history, the performance of a physical examination, the interpretation of laboratory tests, even the definitions of diseases ... are surprisingly inexact. Data is defined, collected, and interpreted with variability and inaccuracy, which falls far short of the standards which engineers do expect from most data. Moreover, standards might be interpreted variably by different medical doctors, different hospitals, different medical schools, and different medical cultures. In particular the last issue is of extreme importance: every clinic, every department, every hospital has its own established standards, and if you are a patient transferred from one to another hospital it is like changing between "different worlds". Organizational culture and communication has actually an important influence on the implementation of IT in Hospitals (Xie et al., 2013).



In order to provide information a patient must first become a patient, so the patient must perceive himself as sick, but patients have different thresholds for the definition of sickness or healthy respectively. The typical patient-doctor dialog uses two types of data: 1) expressed by the patient or the doctor; 2) directly obtained from the patient by the doctor. This is important, because as we will learn in →Lecture 7, the data expressed passes a complex series of perceptive, emotional and cognitive "filters", thus subject to distortion. The types of medical data differentiated by Komaroff (1979) includes expressed data: Verbally expressed objective (past medical history, current illness description, statements, etc.) and verbally expressed subjective (feelings, assumptions, etc.), and Nonverbally expressed (appearance, habitus, mimic, gestures, etc.). The second type is directly obtained data: Elements of physical examination, diagnostic laboratory tests, images, pathognomonic (signs, patterns, etc.).
The big difference between medicine and engineering is, that in medicine a substantial degree of uncertainty may be inevitable; it may not be possible to acquire the needed data, because the measurements cannot be made without destructive consequences for the patient, or of practical limitations, or the length of time required to take adequate measurements. In engineering, given adequate resources, the goal is to reduce uncertainty to a measurably trivial level, and to experimentally demonstrate that the predicted specifications have been met (Komaroff, 1979).



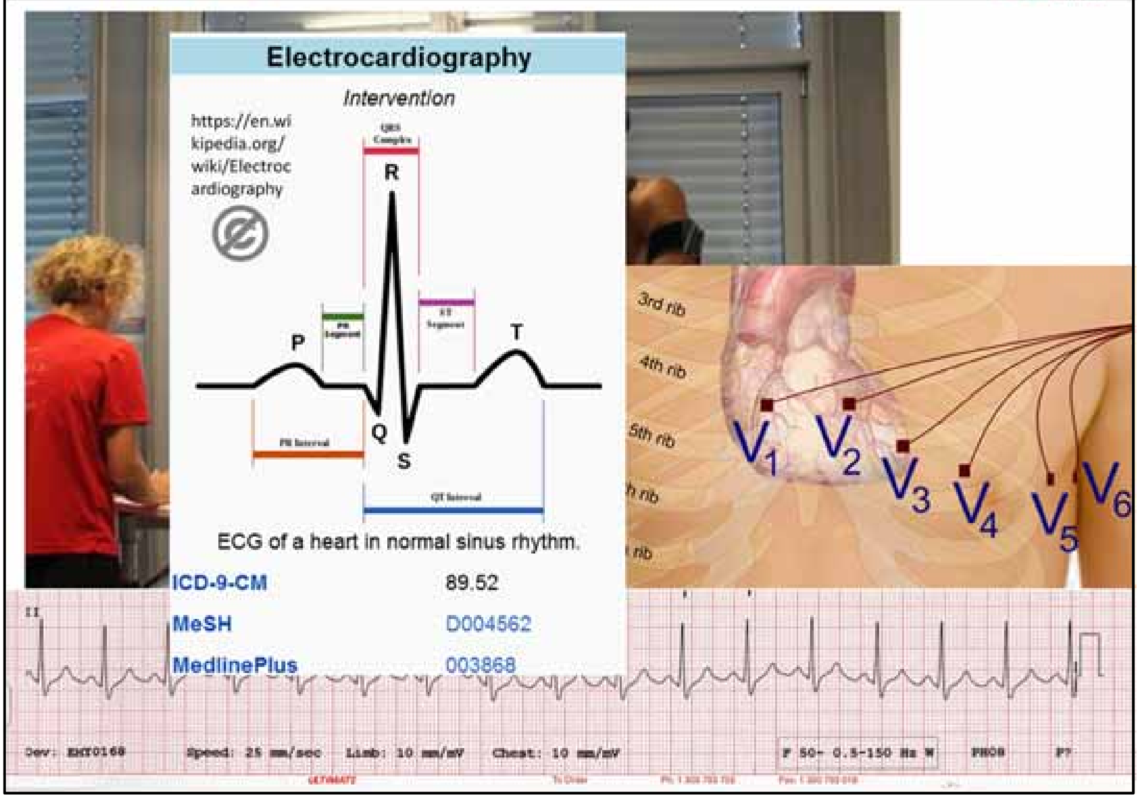
Standardized data shall now ensure that information is interpreted by all users with the same understanding. Moreover, standardized data shall support the reusability of the data, improving the efficiency of healthcare services and avoid errors by reducing duplicated efforts in data entry;
Data standardization refers to:
a) the data content;
b) the terminologies that are used to represent the data;
c) how data is exchanged; and
iv) how knowledge, e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system (refer to IOM).
Elements for sharing require standardization of identification, record structure, terminology, messaging, privacy etc.
The most used standardized data set to date is the International Classification of Diseases (ICD), which was first adopted in 1900 for collecting statistics (Ahmadian et al., 2011). Ahmadian, L., Van Engen-Verheul, M., Bakhshi-Raiez, F., Peek, N., Cornet, R. & De Keizer, N. F. 2011. The role of standardized data and terminological systems in computerized clinical decision support systems: Literature review and survey. International Journal of Medical Informatics, 80, (2), 81-93.
Let us look first at possibly the most difficult example: linguistics in Slide 3-5 and then a manageable example from the recording of an Electrocardiogram (ECG) in Slide 3-6 to emphasize why interoperability is important.



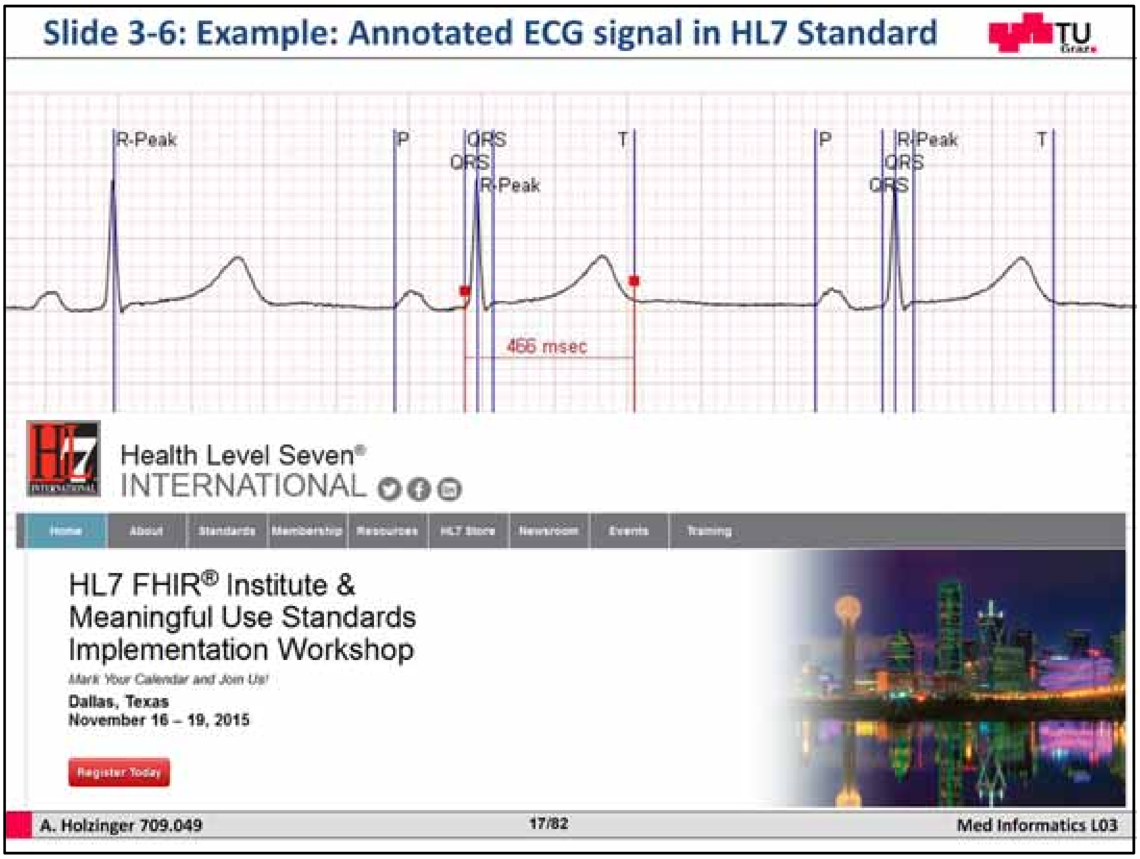
Although we live in a "multimedia age" and some scientists foresee a world without text, in the hospital the major medical documentation is only available in text format and the amount of this unstructured data is immensely increasing (Holzinger et al., 2008), (Holzinger et al., 2013). **Text is the written form of natural language.** Representation of natural language data presents many major challenges. It is difficult to automatically interpret even well-edited texts as well as a native speaking reader would understand it. However, there have been advances in natural language processing (NLP), e.g. the so-called "bag of words" methods: in which a document is treated as a collection of words occurring with some frequency; this works because they do not obscure this inherent meaning when presented to the analyst.

The first mechanized methods were developed by Salton (1968) for information retrieval. Salton's work on identifying salient terms in a corpus, indexing, and constructing high-dimensional signature vectors that represent a corpus' topics or articles remains key to most of the current tools for analyzing big text data (Salton, Wong & Yang, 1975). A challenge is in mapping back the high dimensional vectors into 2D (or 3D) representations to support visualizations that end-users may understand and work on. In addition to Salton's work, centuries of general linguistic study of language provide a foundation for the computer-based analysis of language. The general structure of language provides a framework for the eventual reduction of text to its meaningful logical form for computer-based analysis. While computer-based linguistic analysis is not a solved problem, current capabilities provide some reliable results that add **semantic richness** to the "bag of words" approach. Linguistics defines the levels of structure based on analysis across and within languages, and computational linguistics provides the methods for assigning structure to textual data.
As shown in Slide 3-5, the major levels of structure applicable on text are phonological, morphological, syntactic, semantic, and the pragmatic level:
Phonological level deals with the structure of the sounds that convey linguistic content in a language. However, this level of structure applies to writing and sign language as well. It is basically the lowest level containing the elements that distinguish meaning and can be defined physically as a means of linguistic production.
Morphological level of a language is the level at which meaning can be assigned to parts of words and the level that describes how morphemes (the smallest meaning elements of words) are combined to produce such a word.
Syntactic level of structure concerns the structure of the sentence, i.e., the categories of words and the order in which they are assembled to form a grammatical sentence. The categories used in syntax are known as parts of speech. The main parts of speech are nouns and verbs. Verbs govern the roles that the nouns in the sentence can play, and the ordering and/or case marking of nouns determine their roles.
Semantic level of structure of the sentence is computationally defined to be the level of representation supporting inferencing and other logical operations. WordNet is the preeminent lexicon structured along psycholinguistic principles (Miller, 1998). The utility of WordNet for computational linguistics has been immeasurable. It contains an ontology of the words of English and allows the user to find synonyms, antonyms, hypernyms (more general terms), and hyponyms (more specific terms). It also distinguishes the sense of the words. Other languages have WordNets developed for them and the senses of the words have been linked cross-lingually for use in sense disambiguation within and across languages (see EuroWordNet at http://www.illc.uva.nl/EuroWordNet) (Thomas & Cook, 2005).

Example: ECG



As an example we take a ECG: Electrocardiography (ECG in British English and EKG in American English is the process of recording the electrical activity of the heart over a period of time using electrodes placed on a patient's body. These electrodes detect the tiny electrical changes on the skin that arise from the heart muscle depolarizing during each heartbeat.

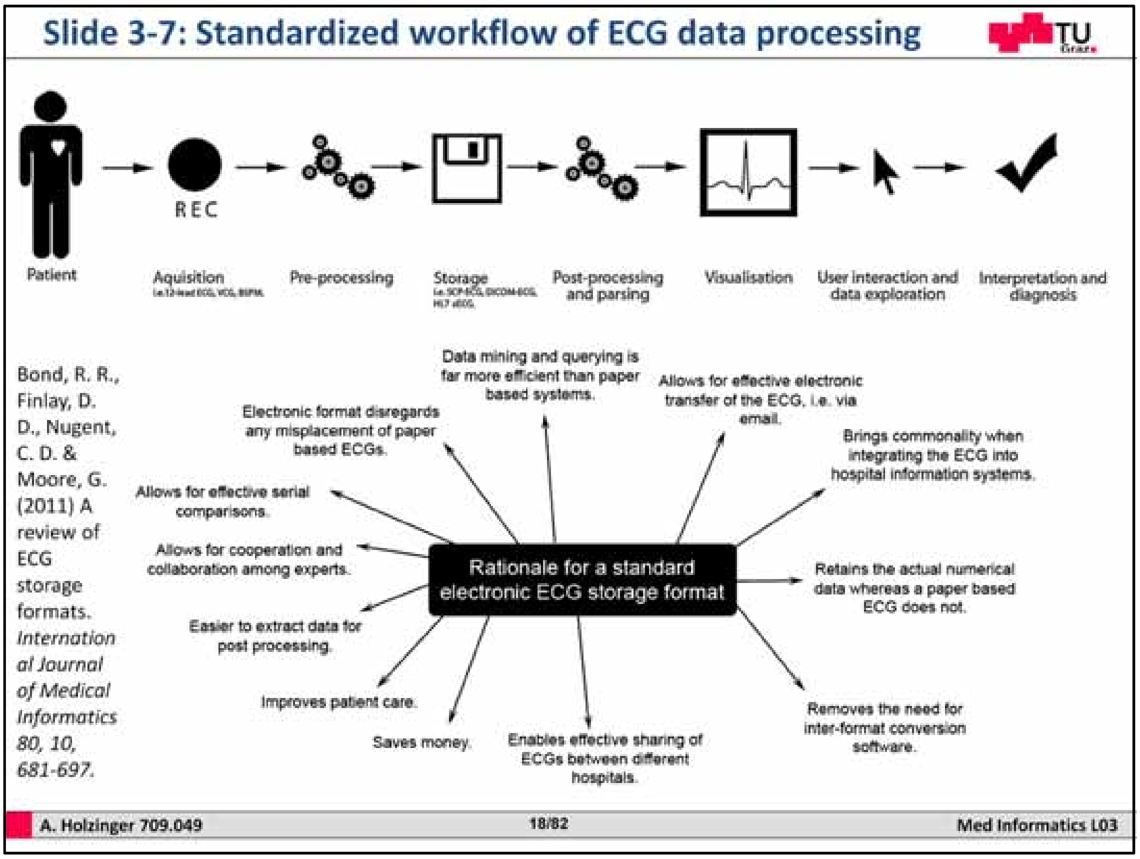


After this complex example, let us look on the recording of an Electrocardiogram (ECG) to explain why interoperability is so important.

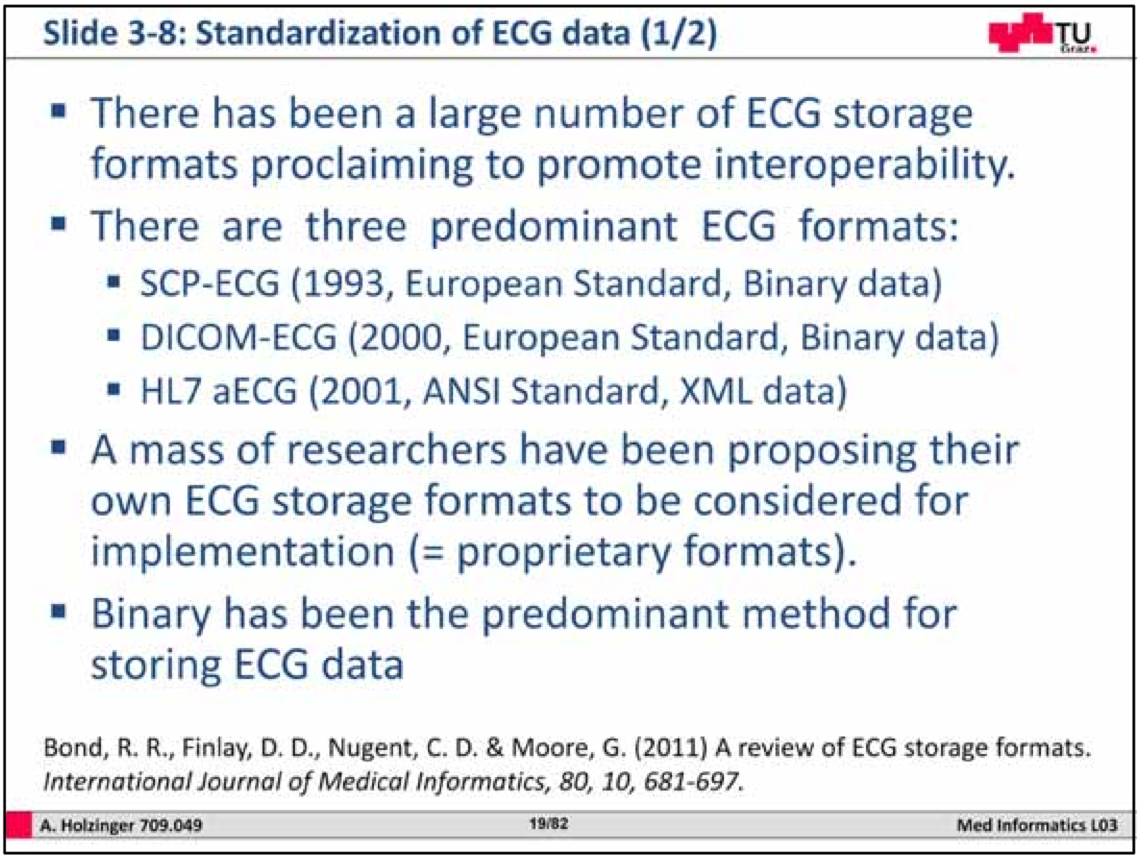
Electrocardiograms are used to measure the rate and regularity of heartbeats, as well as the size and position of the heart chambers.

The importance of creating a standardized ECG data format is reinforced with the increasing demand for interoperability, which is concerned with the coherent exchange of clinical data within and between heterogeneous Hospital Information Systems (HIS). The aim is to facilitate the exchange of medical data, ideally on a global scale. With regards to the ECG, interoperability can only be achieved following the creation of a standardized ECG storage format.

http://www.hl7.org/
HL7 and its members provide a framework (and related standards) for the exchange, integration, sharing, and retrieval of electronic health information. These standards define how information is packaged and communicated from one party to another, setting the language, structure and data types required for seamless integration between systems. HL7 standards support clinical practice and the management, delivery, and evaluation of health services, and are recognized as the most commonly used in the world.

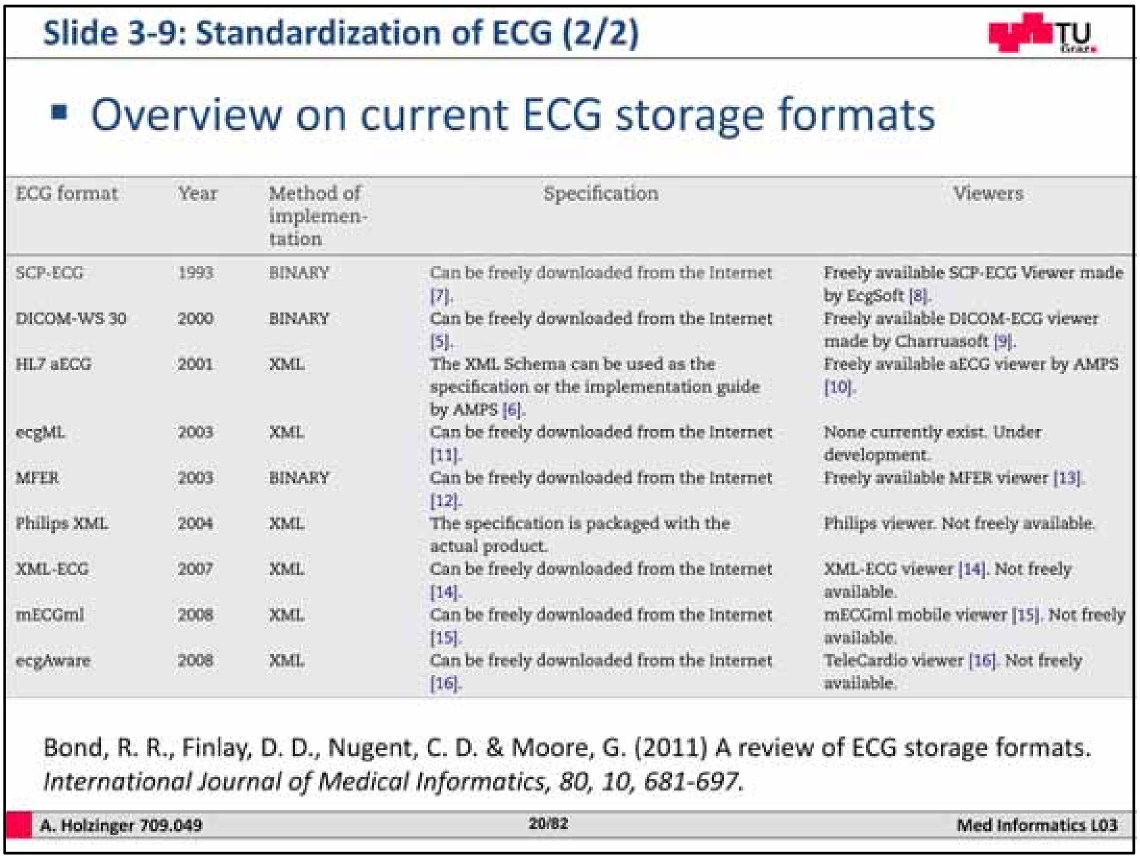


The aim of the standardized data is that the interpretation and diagnosis can be done technically trans-cultural and inter-subjective.
Above we see the typical procedure in the recording and management of an ECG. The importance of creating a standardized ECG storage format is reinforced with the increasing demand for interoperability.
Interoperability is concerned with the coherent exchange of clinical documents within and between heterogeneous Hospital Information Systems. This concept is important since its ultimate aim is to facilitate the exchange of medical data on a global scale. However, it is estimated that this could take still 20 years to achieve effective interoperability in Europe.
Below we can see the rationales for creating a standard electronic ECG storage format (Bond et al., 2011), e.g. the possibility of the application of data mining algorithms on ECGs, or the easy exchange with other health providers.



HL = Health Level
SCP = Standard Communications Protocol
DICOM = Digital Imaging and Communications in Medicine

A huge problem was that so many researchers had proposed their own ECG storage formats and there are many formats proclaiming to promote interoperability, with three predominant ones:
1) SCP-ECG - developed in 1993, stores in binary form, and has been the official European standard for the storage and transmission of ECGs since 2005. In July 2002, the SCP-ECG format became the promotion of a European funded consortium called **OpenECG,** which is a body of at least 464 members who are dedicated to the interoperability in digital electrocardiography. Advantage: small file sizes; Disadvantage: lacking human readability and large number of optional features.
2) DICOM-ECG – originally a image standard called ACR-NEMA in 1985 – it became European standard in 1995, and although the DICOM format was originally created to store and transmit radiographic images, it can now support all diagnostic modalities. As a result NEMA has been extending the DICOM format by developing and publishing supplements. In the year 2000, DICOM-WS 30 was introduced to support the storage of raw medical waveforms, which in effect stores actual sample values as opposed to storing raster images. This supplement has enabled the DICOM format to store various waveform datasets including blood pressure, audio and ECG. Advantage: The power of this format (can display and work as a PACS system, e.g. an ECG and an angiogram at the same time); Disadvantage: Binary based, therefore lacks human readability; too complex.
3) HL7 aECG (annotated ECG) – In November 2001 released by the FDA as a Health Level 7 standard – the first which used XML. Advantage: XML; Disadvantage: verbosity of XML files, consequently large file sizes, uses a lot of definable metadata. For more details please refer to (Bond et al., 2011).



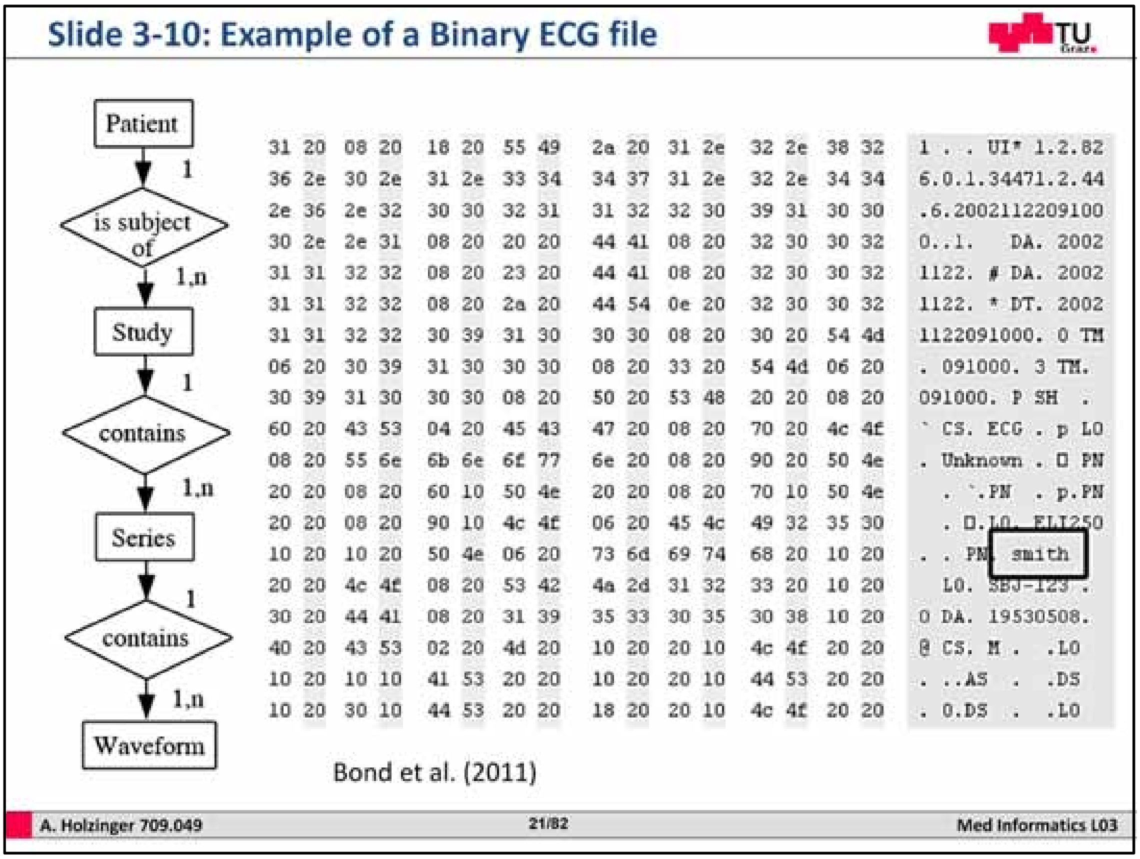
ECG format	Year	Method of implementation	Specification	Viewers
SCP-ECG	1993	BINARY	Can be freely downloaded from the Internet [7].	Freely available SCP-ECG Viewer made by EcgSoft [8].
DICOM-WS 30	2000	BINARY	Can be freely downloaded from the Internet [5].	Freely available DICOM-ECG viewer made by Charruasoft [9].
HL7 aECG	2001	XML	The XML Schema can be used as the specification or the implementation guide by AMPS [6].	Freely available aECG viewer by AMPS [10].
ecgML	2003	XML	Can be freely downloaded from the Internet [11].	None currently exist. Under development.
MFER	2003	BINARY	Can be freely downloaded from the Internet [12].	Freely available MFER viewer [13].
Philips XML	2004	XML	The specification is packaged with the actual product.	Philips viewer. Not freely available.
XML-ECG	2007	XML	Can be freely downloaded from the Internet [14].	XML-ECG viewer [14]. Not freely available.
mECGml	2008	XML	Can be freely downloaded from the Internet [15].	mECGml mobile viewer [15]. Not freely available.
ecgAware	2008	XML	Can be freely downloaded from the Internet [16].	TeleCardio viewer [16]. Not freely available.

This slide shows an overview of some important ECG storage formats, for details please refer to (Bond et al., 2011).

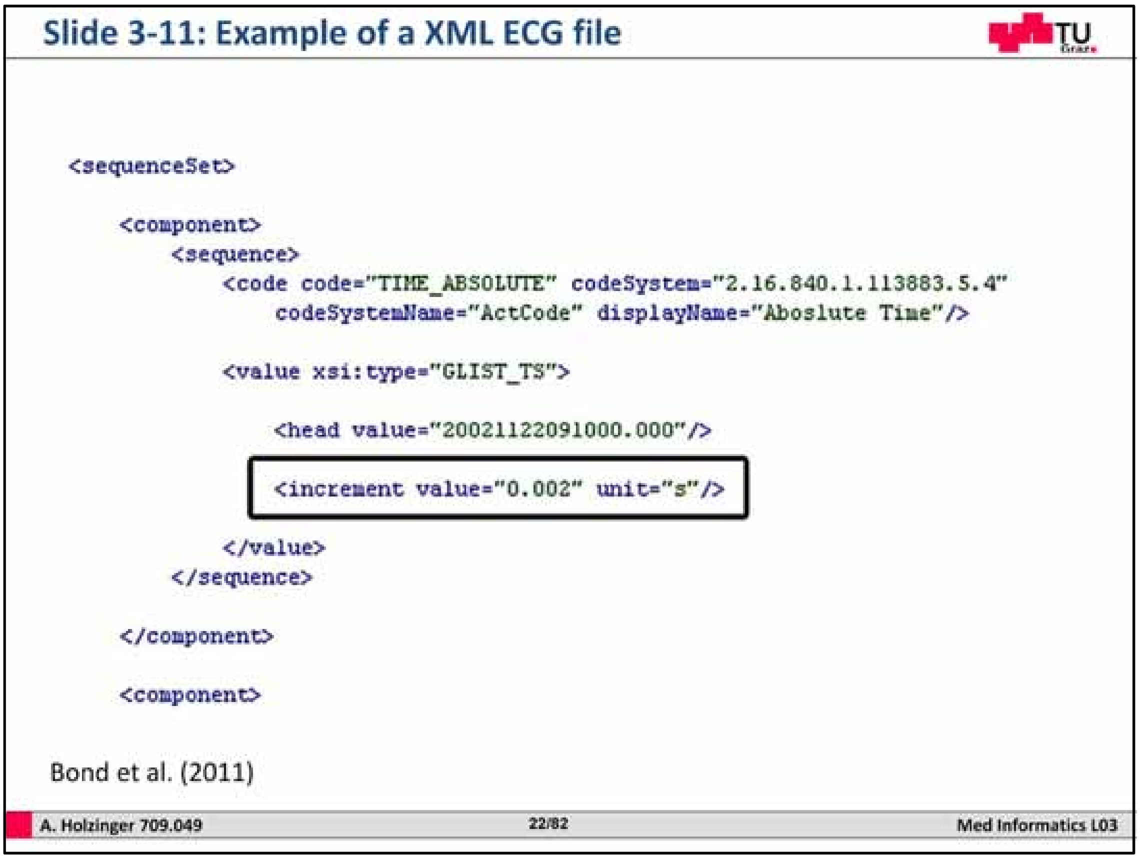
Please remember:

A **binary file** (Binärdatei) contains patterns of bits (Bitmuster) but is not text, although it may contain parts that can be interpreted as text. The disadvantage is that it is not human readable.

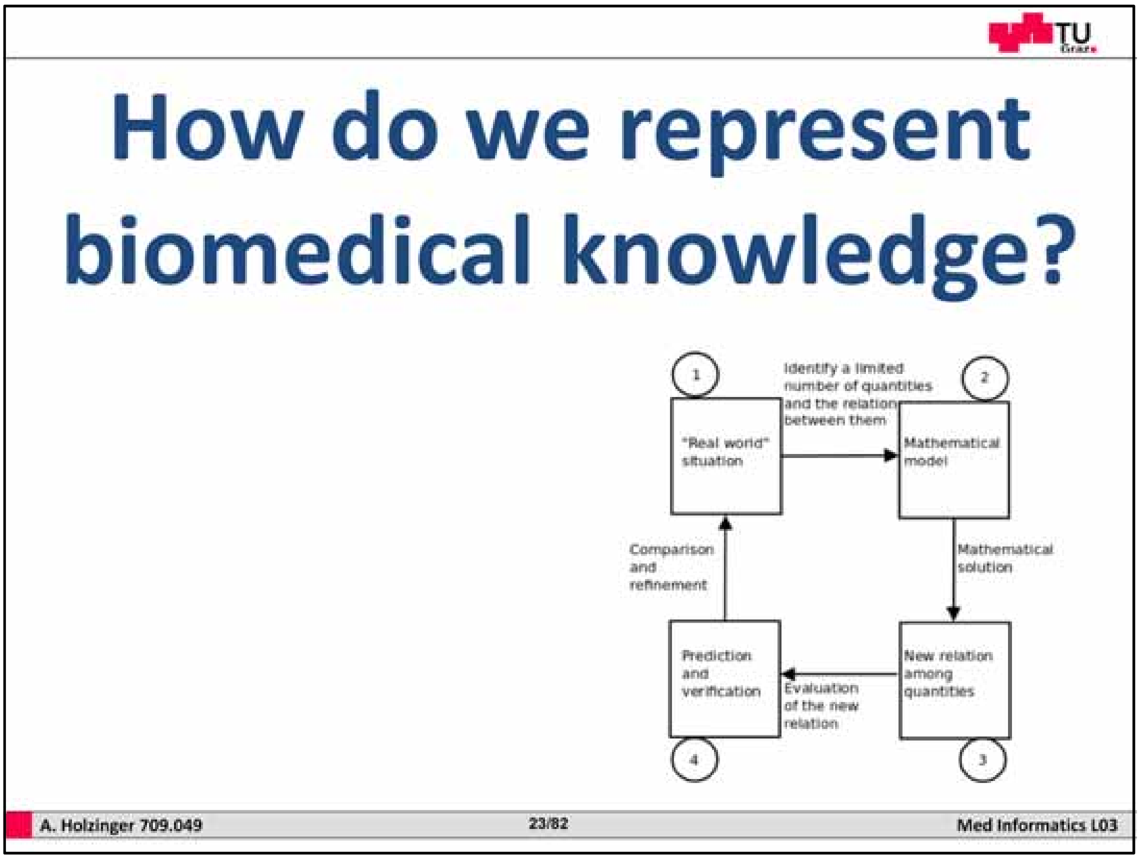
A **XML file** is a string of characters and every legal Unicode character may appear in an XML file, the advantage is that most of the data is human readable.



To demonstrate the big difference of the two file formats, Slide 3-10 shows an example file in Binary, and Slide 3-11 an example file in XML.



Here we see a typical example of an aECG file indicating the increment element which defines the interval in seconds between each sample. The value 0.002 s indicates that there is a two millisecond gap between each sample. This, in effect would be the frequency equivalent of 500 Hz (Bond et al., 2011).

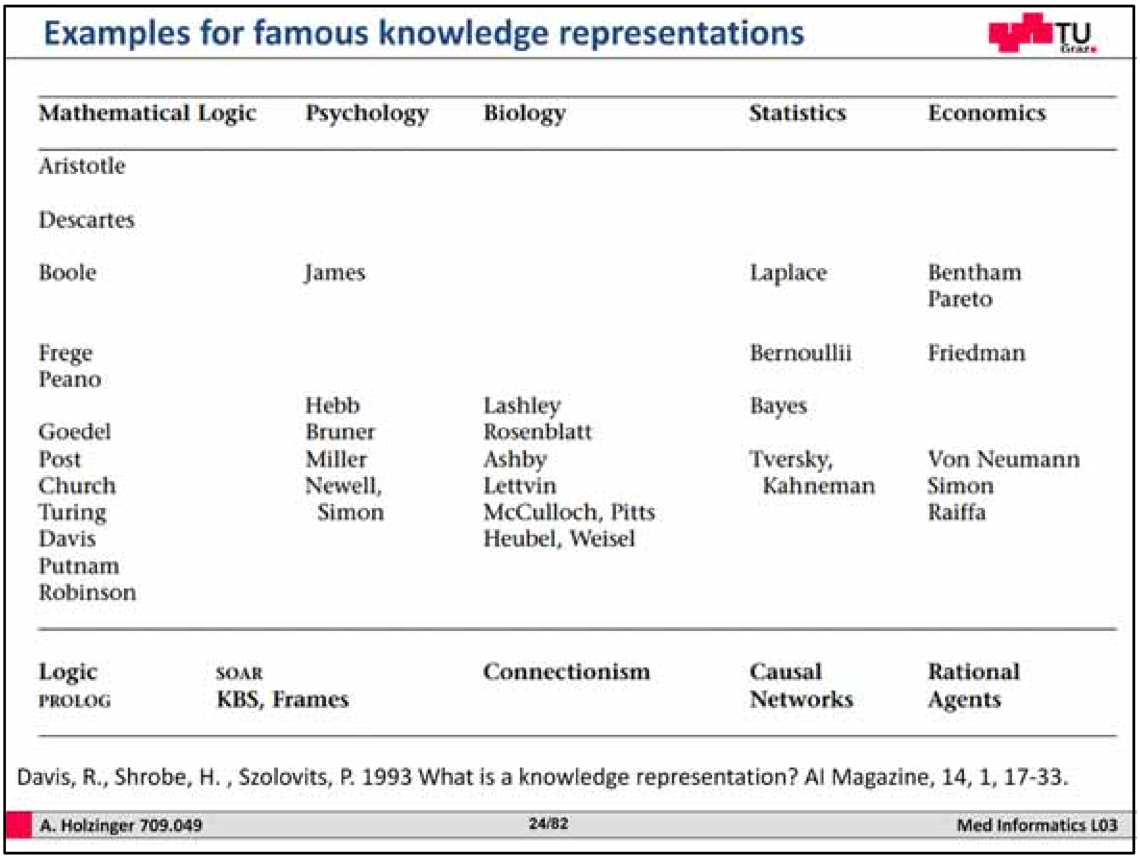


What does modelling mean?

Knowledge modeling is a process of creating a computer interpretable model of knowledge or standard specifications about a kind of process and/or about a kind of facility or product. The resulting knowledge model can only be computer interpretable when it is expressed in some knowledge representation language or data structure that enables the knowledge to be interpreted by software and to be stored in a database or data exchange file

Knowledge representation and reasoning (KR) is the field of artificial intelligence (AI) dedicated to representing information about the world in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a natural language. Knowledge representation incorporates findings from psychology about how humans solve problems and represent knowledge in order to design formalisms that will make complex systems easier to design and build. Knowledge representation and reasoning also incorporates findings from logic to automate various kinds of reasoning, such as the application of rules or the relations of sets and subsets.

The earliest work in computerized knowledge representation was focused on general problem solvers such as the General Problem Solver (GPS) system developed by Allen Newell and Herbert A. Simon in 1959. These systems featured data structures for planning and decomposition. The system would begin with a goal. It would then decompose that goal into sub-goals and then set out to construct strategies that could accomplish each subgoal.



Examples for famous knowledge representations

Mathematical Logic	Psychology	Biology	Statistics	Economics
Aristotle				
Descartes				
Boole	James		Laplace	Bentham Pareto
Frege Peano			Bernoullii	Friedman
	Hebb	Lashley	Bayes	
Goedel Post Church Turing Davis Putnam Robinson	Bruner Miller Newell, Simon	Rosenblatt Ashby Lettvin McCulloch, Pitts Heubel, Weisel	Tversky, Kahneman	Von Neumann Simon Raiffa
Logic PROLOG	SOAR **KBS, Frames**	**Connectionism**	**Causal Networks**	**Rational Agents**

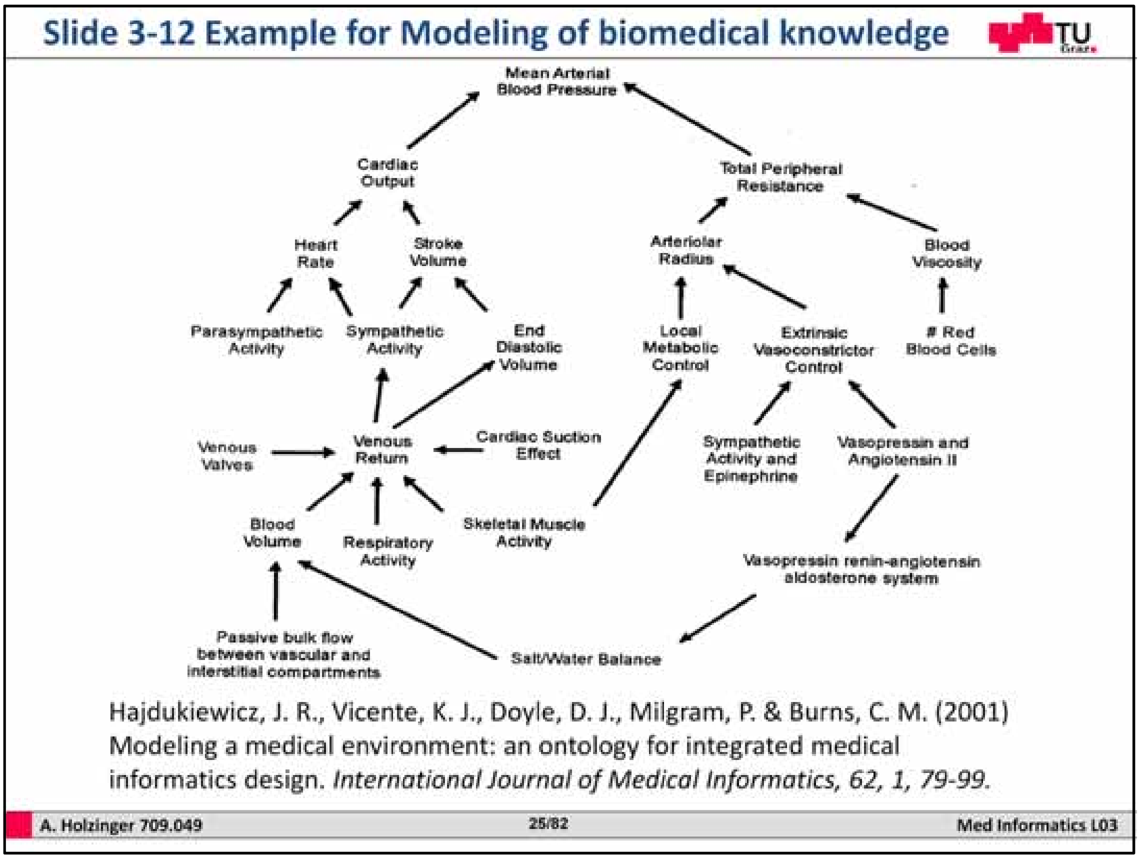
Davis, R., Shrobe, H. , Szolovits, P. 1993 What is a knowledge representation? AI Magazine, 14, 1, 17-33.

A. Holzinger 709.049 24/82 Med Informatics L03

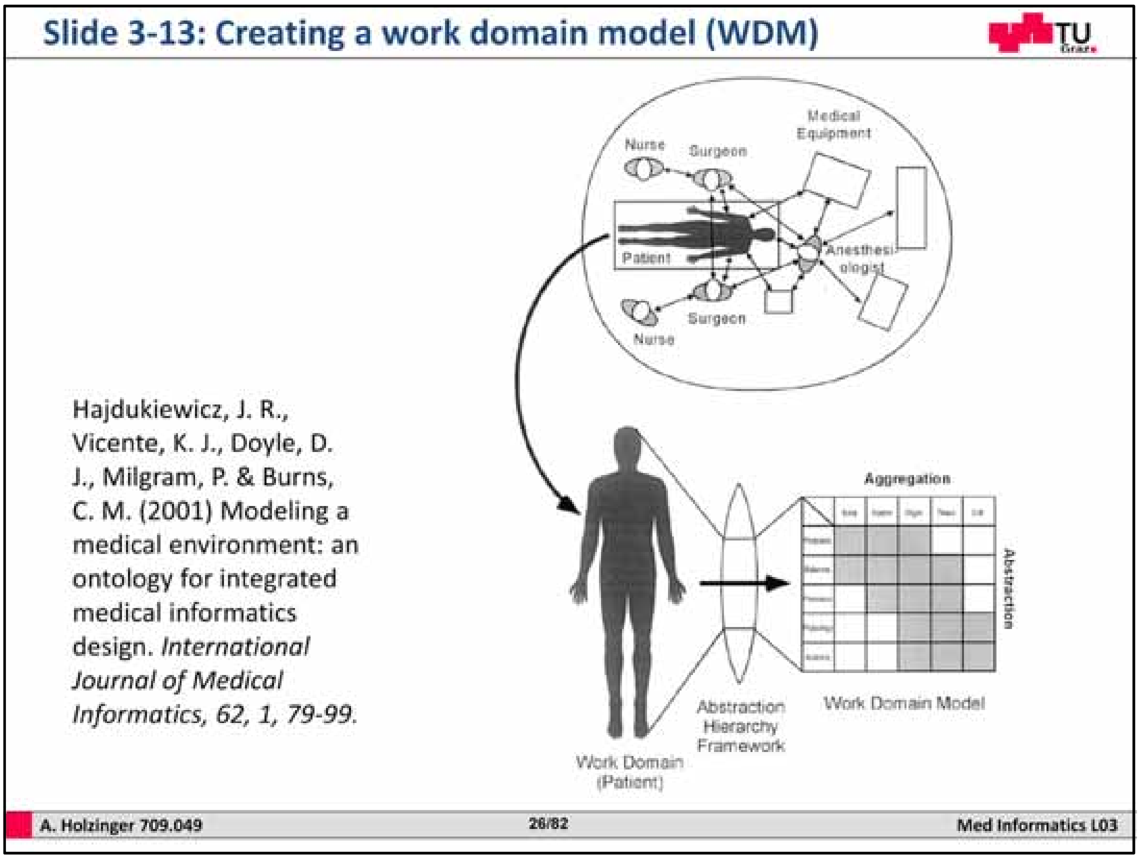
http://groups.csail.mit.edu/medg/ftp/psz/k-rep.html

Inference means any way to get new expressions from old ones.

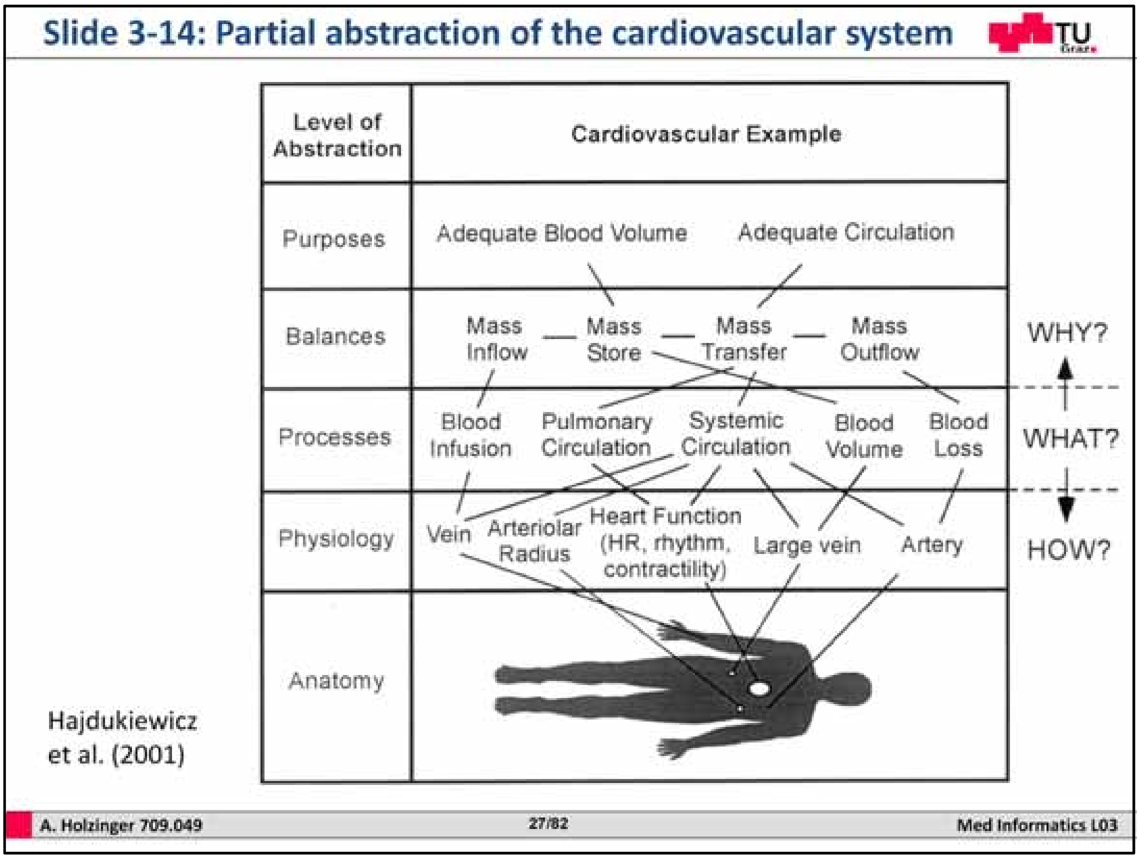
A Knowledge Representation is:
1) a Surrogate
2) a Set of Ontological Commitments
Reminder: A Knowledge Representation Is Not a Data Structure: A semantic net, for example, is a representation, but a graph is a data structure.
3) a Fragmentary Theory of Intelligent Reasoning
4) a Medium for Efficient Computation
5) a Medium of Human Expression



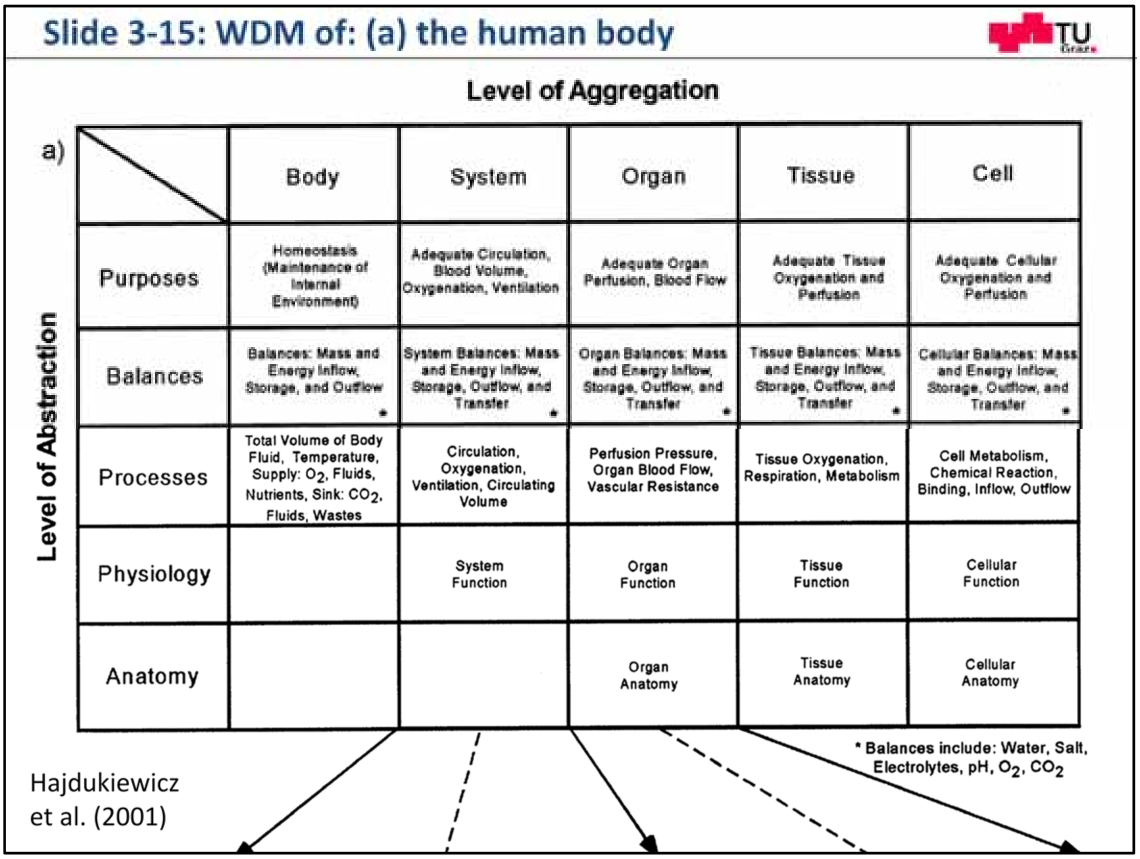
Medical environments have enormous complexity and poses high demands on medical professionals. Here we see an example of a traditional modeling approach for medical reasoning used as a basis for developing decision support systems. Such models may be faithful to what is known about biomedical knowledge, but they have (serious!) limitations for human problem solving, especially in unanticipated situations. This example shows the physiological factors and relations affecting mean arterial blood pressure (Hajdukiewicz et al., 2001).



This Slide illustrates the process of generating a WDM of the patient (i.e. the human body) in an operating room (OR). The OR consists of a team of medical personnel (2 nurses, 2 surgeons, 1 anesthesiologist) who interact with each other, the patient, and with medical equipment to perform a surgical procedure.
We define "work domain" as an object in this environment that is controlled and, due to its complexity and purpose, can require problem solving by the medical personnel. A work domain could be the patient himself or a complex medical device (e.g. anesthesia workstation). In the example by Hajdukiewicz et al. the work domain is the patient. The patient WDM is divided into different levels of abstraction (Abstraction Hierarchy, AH) and aggregation (part–whole hierarchy, PWH). Each "cell" (another meaning of the word "cell" ;-) in this patient WDM matrix defines a complete and different causal representation of the same patient work domain, uniquely defined by the particular levels of abstraction and aggregation.



If we now "zoom-in" we see the structural means–ends links between the different levels of abstraction for parts of the patient cardiovascular system. The lower levels include the cardiac and circulatory functions necessary to support the higher-level purposes of adequate circulation and blood volume; the higher levels provide reasons for lower level functions. Here the problem solving can occur by shifting the mental focus across these levels of abstraction. Information will be required from the Abstraction Hierarchy (AH) level currently in the practitioner's mental focus, including the functional structure, state, and what needs to be controlled (i.e. the What?). For example, the current task may be to control systemic circulation. Information is also required from the AH level above, which indicates the reason of the control decision (i.e. the Why?). In this Slide the reasons for controlling systemic circulation are to support the functions of mass transfer and balance to the organs. Finally, information is required from the AH level below, which indicates the physiological resources available for implementing the decision (i.e. the How?) (Hajdukiewicz et al., 2001).



Slide 3-15: WDM of: (a) the human body

Level of Aggregation

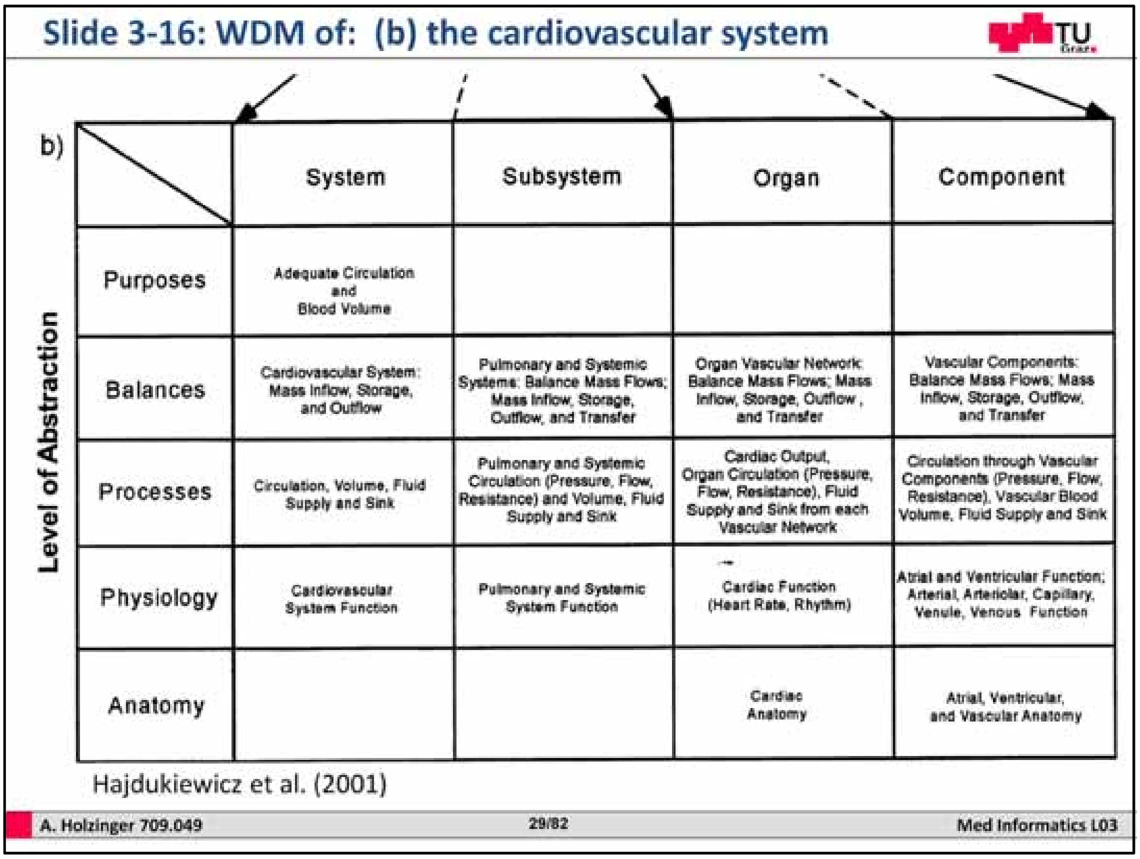
a)

Level of Abstraction	Body	System	Organ	Tissue	Cell
Purposes	Homeostasis (Maintenance of Internal Environment)	Adequate Circulation, Blood Volume, Oxygenation, Ventilation	Adequate Organ Perfusion, Blood Flow	Adequate Tissue Oxygenation and Perfusion	Adequate Cellular Oxygenation and Perfusion
Balances	Balances: Mass and Energy Inflow, Storage, and Outflow *	System Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *	Organ Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *	Tissue Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *	Cellular Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *
Processes	Total Volume of Body Fluid, Temperature, Supply: O_2, Fluids, Nutrients, Sink: CO_2, Fluids, Wastes	Circulation, Oxygenation, Ventilation, Circulating Volume	Perfusion Pressure, Organ Blood Flow, Vascular Resistance	Tissue Oxygenation, Respiration, Metabolism	Cell Metabolism, Chemical Reaction, Binding, Inflow, Outflow
Physiology		System Function	Organ Function	Tissue Function	Cellular Function
Anatomy			Organ Anatomy	Tissue Anatomy	Cellular Anatomy

* Balances include: Water, Salt, Electrolytes, pH, O_2, CO_2

Hajdukiewicz et al. (2001)

A further "zoom-in" into each cell, where we find a model consisting of different objects or functions connected by causal relations – further detailed in Slide 3-16.



Slide 3-16: WDM of: (b) the cardiovascular system

b)

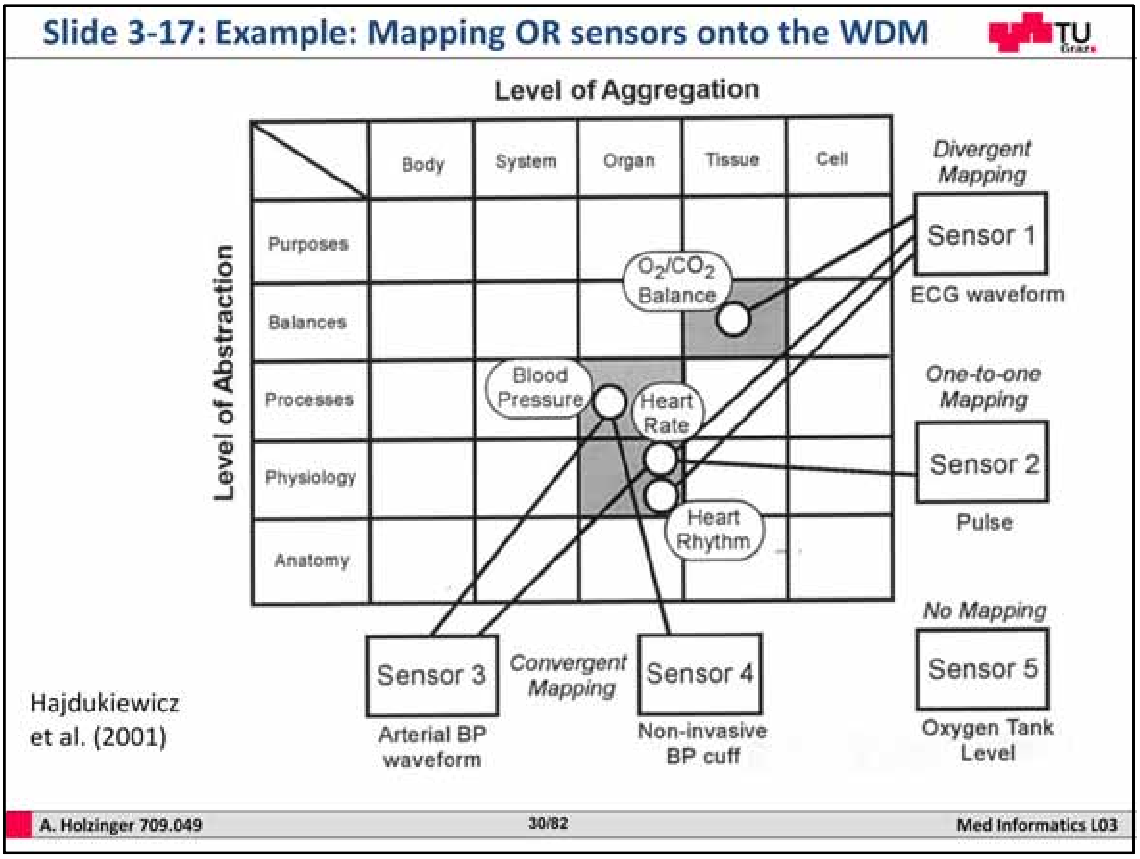
Level of Abstraction

	System	Subsystem	Organ	Component
Purposes	Adequate Circulation and Blood Volume			
Balances	Cardiovascular System: Mass Inflow, Storage, and Outflow	Pulmonary and Systemic Systems: Balance Mass Flows; Mass Inflow, Storage, Outflow, and Transfer	Organ Vascular Network: Balance Mass Flows; Mass Inflow, Storage, Outflow, and Transfer	Vascular Components: Balance Mass Flows; Mass Inflow, Storage, Outflow, and Transfer
Processes	Circulation, Volume, Fluid Supply and Sink	Pulmonary and Systemic Circulation (Pressure, Flow, Resistance) and Volume, Fluid Supply and Sink	Cardiac Output, Organ Circulation (Pressure, Flow, Resistance), Fluid Supply and Sink from each Vascular Network	Circulation through Vascular Components (Pressure, Flow, Resistance), Vascular Blood Volume, Fluid Supply and Sink
Physiology	Cardiovascular System Function	Pulmonary and Systemic System Function	Cardiac Function (Heart Rate, Rhythm)	Atrial and Ventricular Function; Arterial, Arteriolar, Capillary, Venule, Venous Function
Anatomy			Cardiac Anatomy	Atrial, Ventricular, and Vascular Anatomy

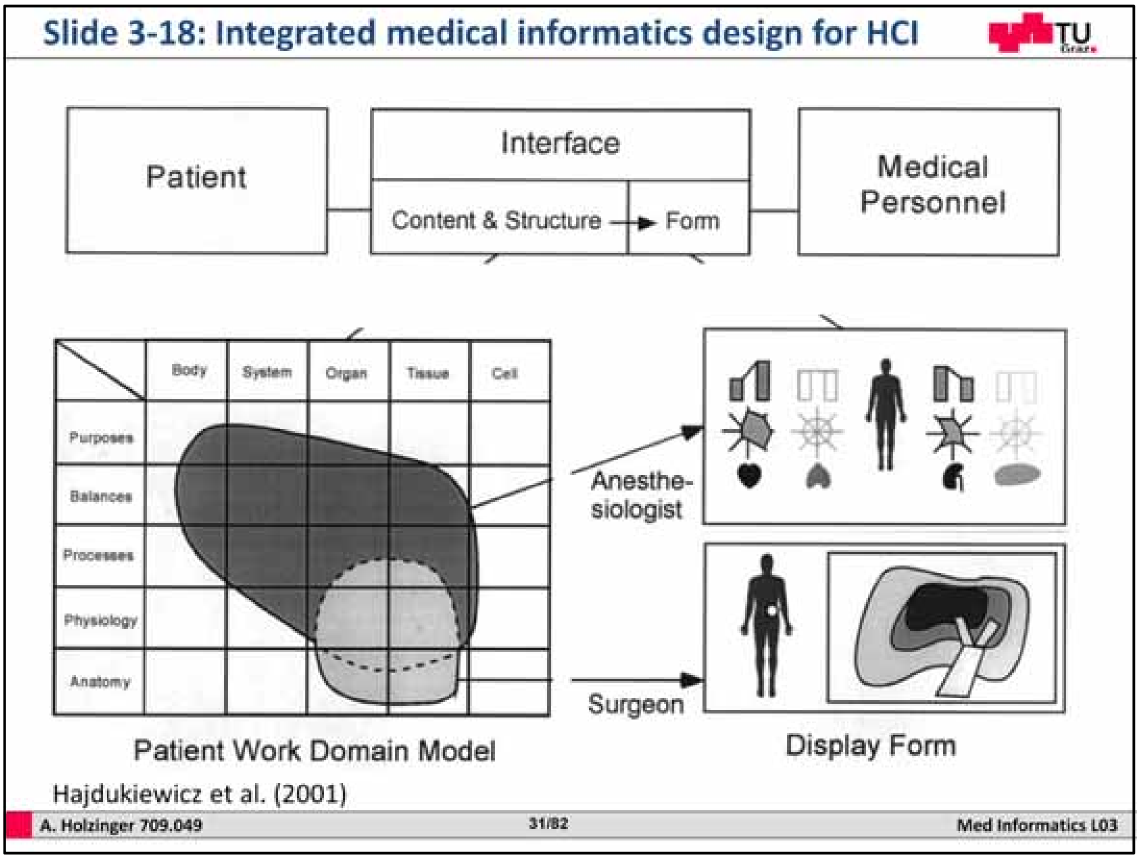
Hajdukiewicz et al. (2001)

A. Holzinger 709.049 29/82 Med Informatics L03

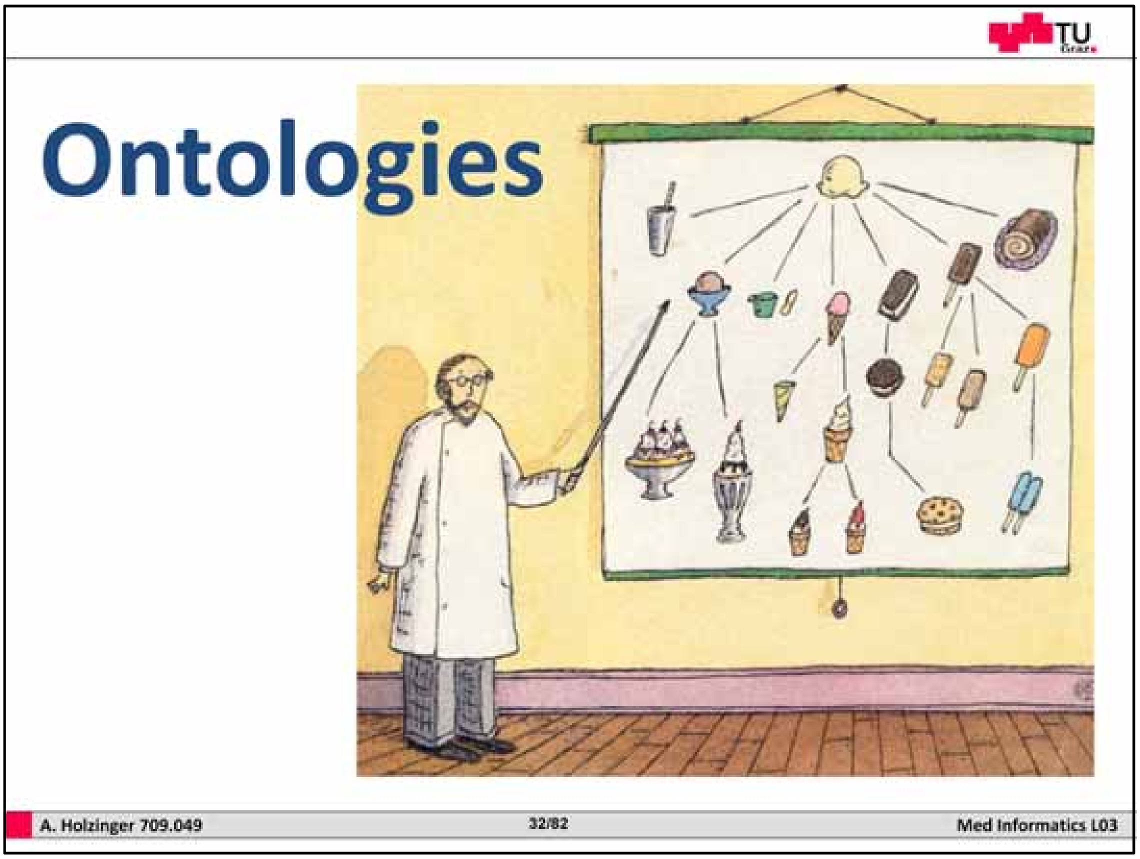
Here we further “zoom-in” and see the causal arrangements for selected parts of the human body that are reasonable to illustrate, given the complexity of the cardiovascular system (i.e. levels of abstraction, balances and processes; levels of aggregation, system, sub-system, organ).



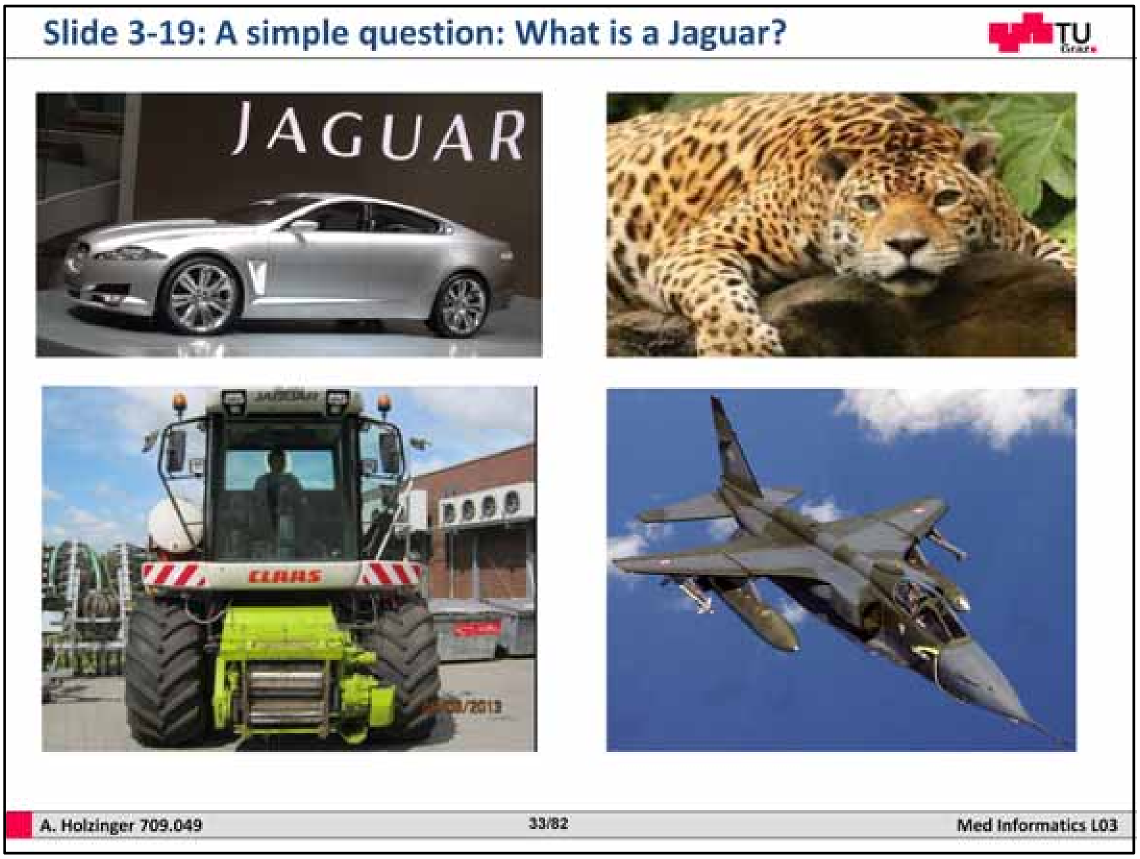
You will now have asked yourself: what is the purpose of such modeling? In Slide 3-17 you see a typical example how useful it can be: We see four types of mapping between the patient WDM and operating room sensors: one-to-one, convergent, divergent, and no mapping. With a one-to-one mapping, one sensor maps onto one patient variable. For example, checking a patient's pulse provides information about heart rate. With convergent mapping (or redundancy), many sensors map onto one patient variable. Practitioners use this method to reduce the high level of uncertainty in measurements from the environment (e.g. artifact, noise, and calibration errors). For example, heart rate can be determined directly from the ECG signal as well as indirectly from other monitor signals (e.g. arterial blood pressure waveforms). With divergent mapping, some sensors provide evidence for many patient variables. The ECG waveform provides evidence for heart rate, heart rhythm, and adequate myocardial oxygenation, etc. Finally, with no mapping, some sensors do not map onto any of those patient variables, e.g. the pressure in an unused oxygen tank (Hajdukiewicz et al., 2001).



Our last example demonstrates the usefulness of WDM for human-computer interaction, in a way that is compatible with how medical practitioners can perform problem solving in the context of a medical environment. Note: The information requirements for surgeons are very different compared with anesthesiologists, although both need the same information from overlapping regions of the patient WDM (Hajdukiewicz et al., 2001).

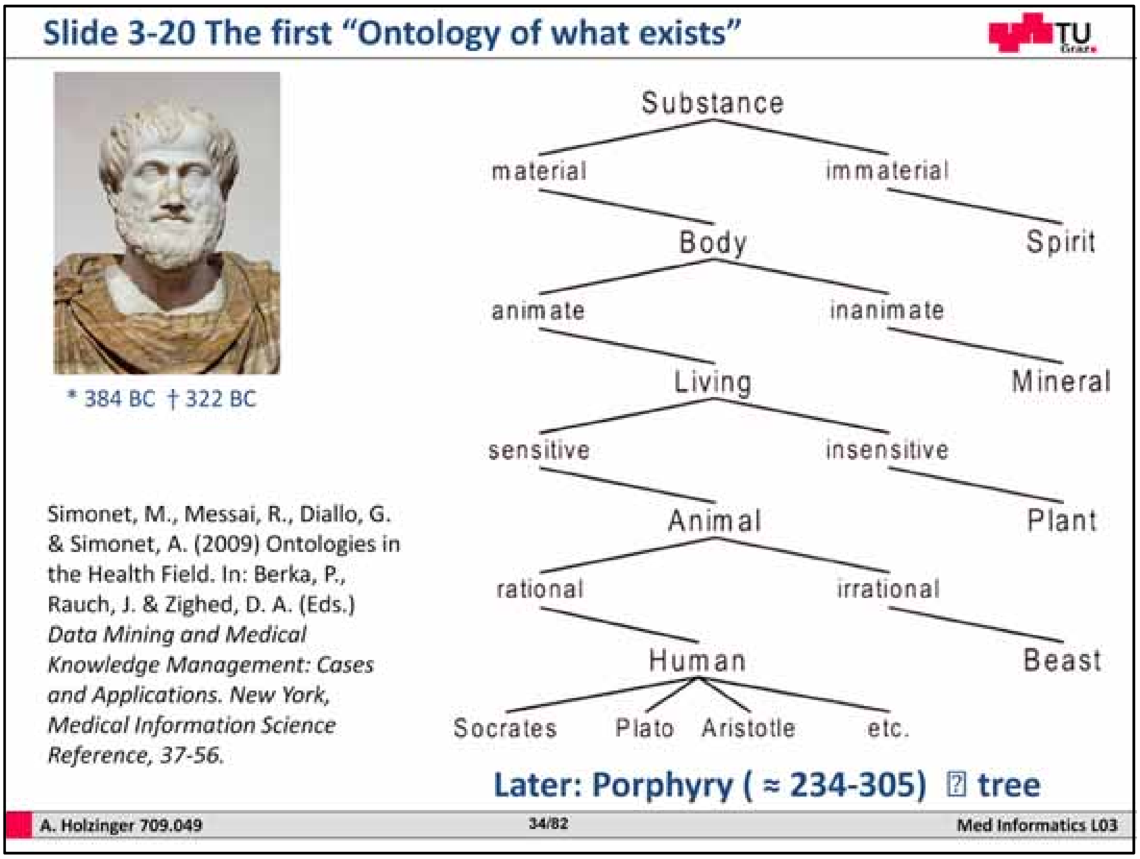


When talking about standardization we immediately touch ontologies. In computer science an ontology represents **formal knowledge as a set of concepts** within a (strictly limited) domain, and the relationships between those concepts. It is similar to what we have seen before, as it can be used for domain modeling.
The most important aspect is that an ontology provides a standardized (shareable) vocabulary, which can then be used to model such a domain.



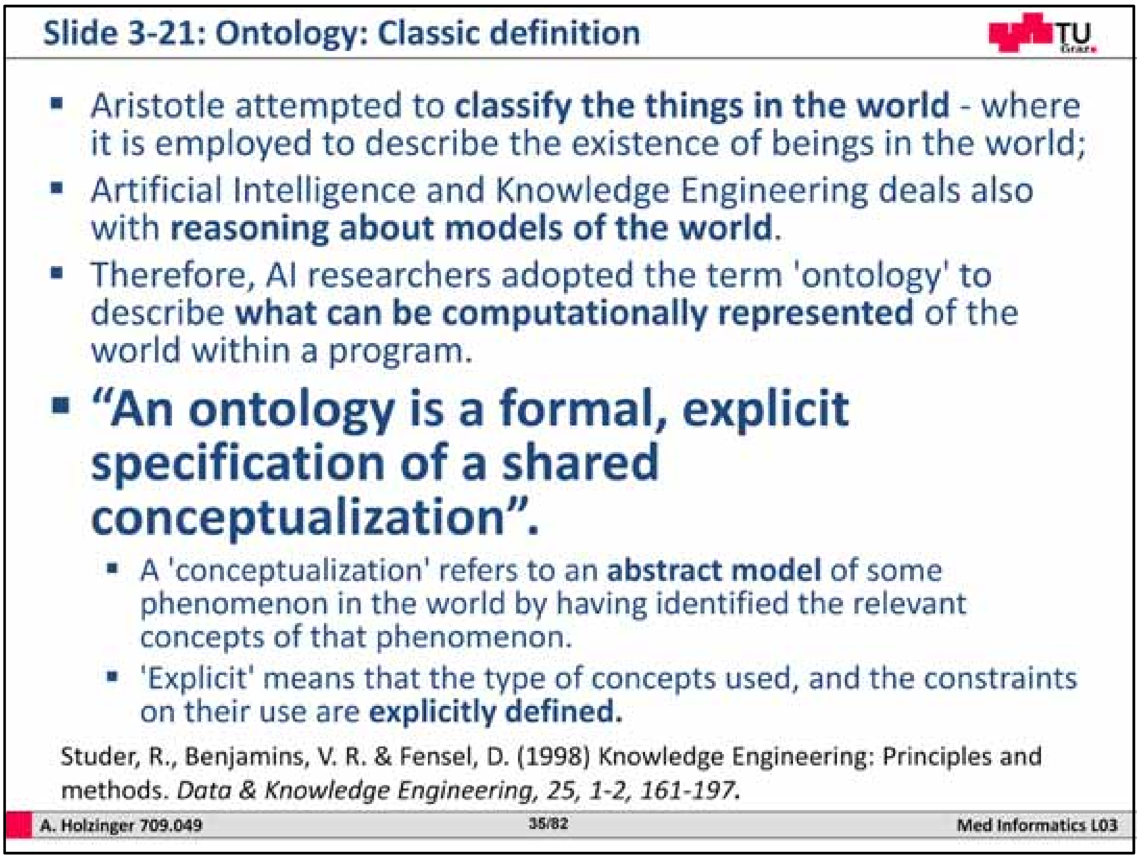
If you put the keyword "Jaguar" into a search engine – you will get different results. The search engine – as a typical Von-Neumann machine – does not know what you are looking for: a sports-car, an animal, a jet plane, or a tractor? Our current computers cannot know in what context you use the word "Jaguar" – so additional (meta-) information is needed. Meta-information is information about information.

A categorization may help – the first known categorization was done by Artistoteles.



An ontology is defined as a theory of reality (in philosophy) or a conceptualization of what exists (in artificial intelligence). In practice, an ontology consists of categories of individuals organized in taxonomies and connected by various other relationships. This is the reason why a graph structure is often used for representing ontologies. In order to be able to assess and enforce the modeling principles for ontologies, we start by defining the following notions: graph structure, taxonomy, and ontology.

Definitions of these notions focus on structural aspects and are not intended to capture all aspects of ontologies (for a formal definition of biological classes and ontological relations, see [12,13]).



In this Slide we see the classic definition: An ontology is a formal, explicit specification of a shared conceptualization (Studer, Benjamins & Fensel, 1998). Ontology IS-A a structured description of a domain in form of concepts ↔ relations; this **IS-A relation** provides a taxonomic skeleton. Other relations reflect the domain semantics and formalize the terminology in this particular domain. The terminology contains the terms and their definitions and usage in a specific context. The knowledge base is the instance classification and concept classification; the classification itself provides the domain terminology (Holzinger, 2000).

Slide 3-22: Ontology: Terminology

- Ontology = a structured description of a domain in form of **concepts ↔ relations**;
- The **IS-A relation** provides a taxonomic skeleton;
- Other relations reflect the **domain semantics**;
- Formalizes the **terminology** in the domain;
- Terminology = terms definition and usage in the specific **context**;
- Knowledge base = **instance classification** and **concept classification**;
- Classification provides the **domain terminology**

...

A. Holzinger 709.049 36/82 Med Informatics L03

Ok, let us review: Ontology is defined as a theory of the reality (in philosophy) or a conceptualization of what exists (in artificial intelligence). In practice, ontologies consist of categories of entities organized in taxonomies and connected by relationships. ARISTOTLE attempted to classify the things in the world, consequently researchers adopted the term 'ontology' to describe what can be computationally represented of the world within machine language (software): An ontology is a formal, explicit specification of a shared conceptualization. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined (Studer, Benjamins & Fensel, 1998).

Slide 3-23: Additionally an ontology may satisfy:

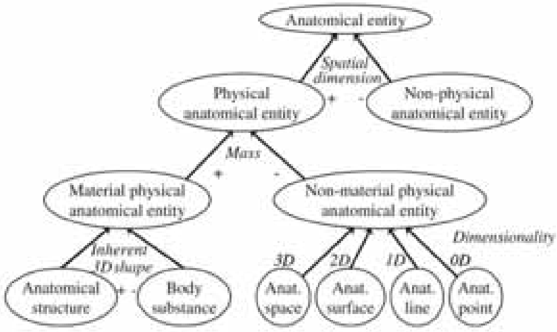
(1) In addition to the IS-A relationship, partitive (meronomic) relationships may hold between concepts, denoted by PART-OF. Every PART-OF relationship is irreflexive, asymmetric and transitive. IS-A and PART-OF are also called hierarchical relationships.

(2) In addition to hierarchical relationships, associative relationships may hold between concepts. Some associative relationships are domain-specific (e.g., the branching relationship between arteries in anatomy and rivers in geography).

(3) Relationships r and r' are inverses if, for every pair of concepts x and y, the relations $\langle x, r, y\rangle$ and $\langle y, r', x\rangle$ hold simultaneously. A symmetric relationship is its own inverse. Inverses of hierarchical relationships are called INVERSE-IS-A and HAS-PART, respectively.

(4) Every non-taxonomic relation of x to z, $\langle x, r, z\rangle$, is either inherited ($\langle y, r, z\rangle$) or refined ($\langle y, r, z'\rangle$ where z' is more specific than z) by every child y of x. In other words, every child y of x has the same properties (z) as it parent or more specific properties (z').

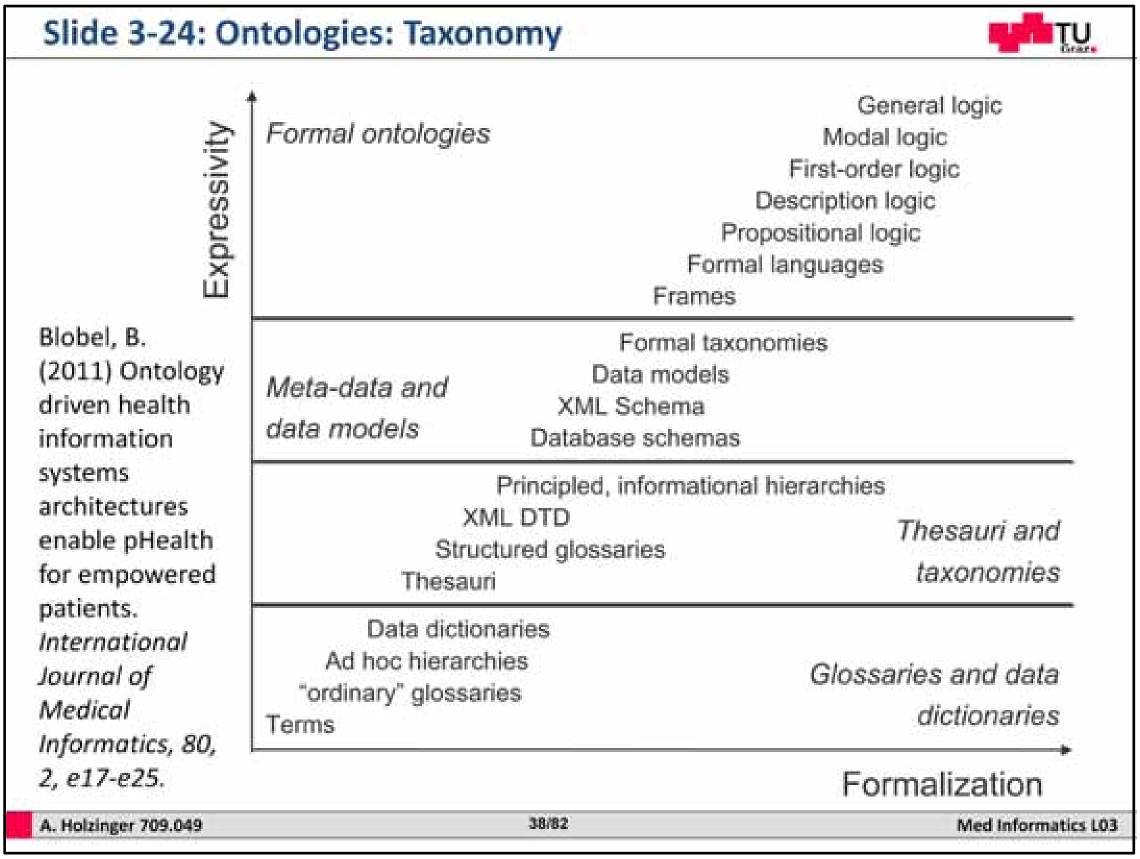
Zhang, S. & Bodenreider, O. 2006. Law and order: Assessing and enforcing compliance with ontological modeling principles in the Foundational Model of Anatomy. *Computers in Biology and Medicine, 36, (7-8), 674-693.*



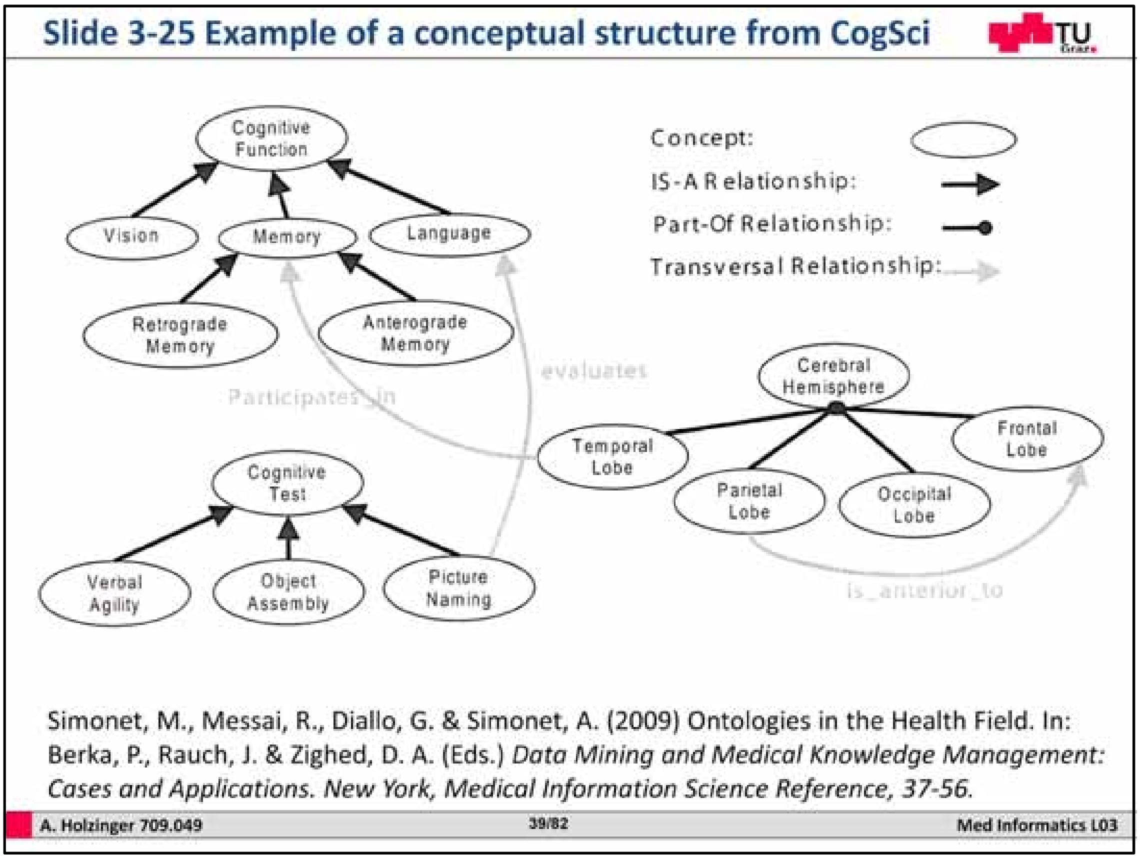
A. Holzinger 709.049 37/82 Med Informatics L03

An *ontology* is composed of at least one taxonomy and may comprise several distinct taxonomies. Concepts across taxonomies do not stand in a taxonomic relation. Concepts in an ontology represent categories of things existing in reality or abstractions generated for classification purposes. Each category or abstraction is represented exactly by one concept (Zhang & Bodenreider, 2006).

Bottom right in Slide 3-23 you can see as an example the top-level of the anatomy taxonomy along with the classification criteria.



In slide 3-24 we see a hierarchy of ontologies regarding formalization on the x-axis and expressivity on the y-axis.
Whereas typical dictionaries are on the left-down corner, first-order logic is on the right-up corner.
Note: Logic programming is a well-known declarative method of knowledge representation based on first-order logic. Logic programming was developed in the early 1970s based on work in automated theorem proving. A logic program consists of a set of rules (Horn clauses), where each rule has the form head body, where head is a logical atom and body is a conjunction of logical atoms. The logical semantics of such a rule is given by the implication body head. The semantics of a pure logic program is completely independent of the order in which its clauses are given, and of the order of the single atoms in each rule body. In PROLOG, the paradigm of logic programming is practically usable. The clause matching and backtracking algorithms at the core of PROLOG are sensitive to the order of the clauses in a program and of the atoms in a rule body. In application areas such as knowledge representation and databases there is a predominant need for full declarativeness, and hence for pure logic programming. In knowledge representation, declarative extensions of pure logic programming, such as negation in rule bodies and disjunction in rule heads, are used to formalize common sense reasoning. In the database context, the query language DATALOG was designed. (Dantsin, Eiter, Gottlob & Voronkov, 2001), (Eiter et al., 2006).



This figure presents an example from cognitive science. The knowledge about the brain domain (aka anatomy-functional ontology) is expressed through semantic relationships between the concepts of the three ontologies; This ontology has been used to support the discovery of relationships between the cognitive function and the anatomical regions of the brain

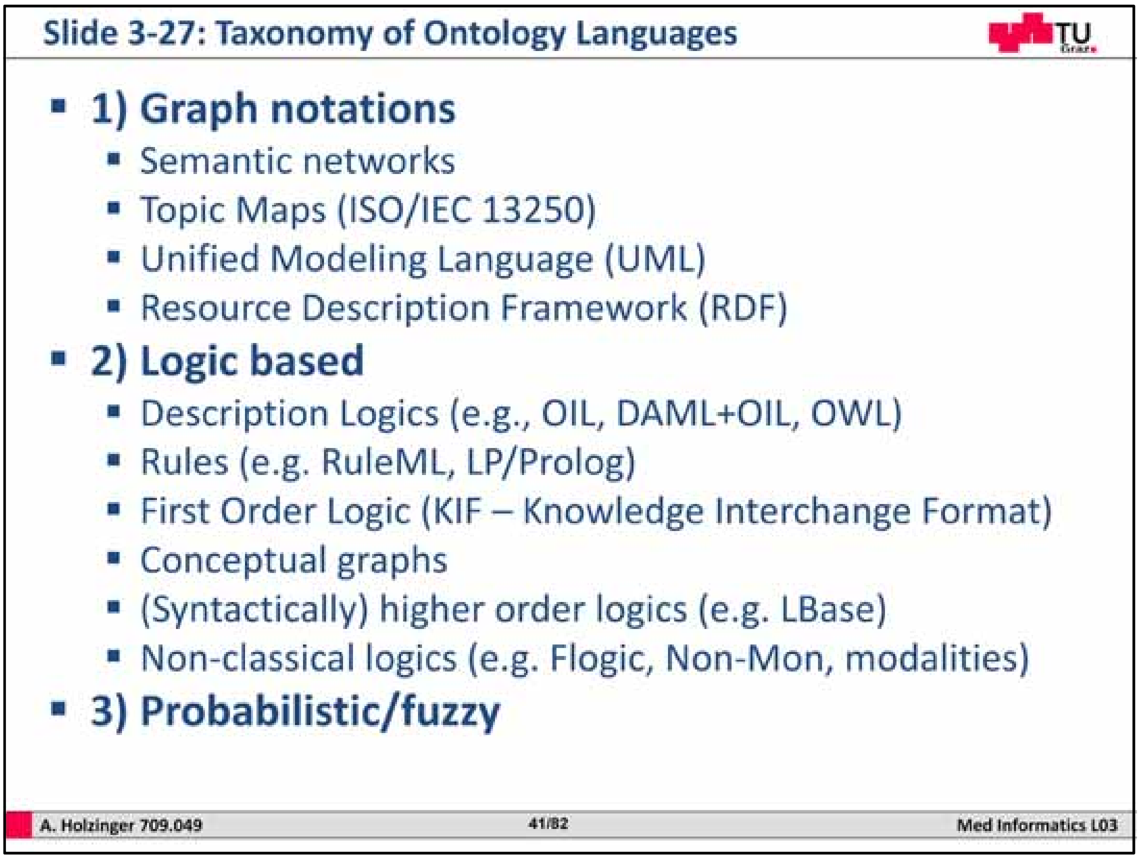
Slide 3-26: Examples of Biomedical Ontologies

Name	Ref.	Scope	# concepts	# concept names Min	Max	Med	Avg	Subs. Hier.	Version / Notes
SNOMED CT	[21]	Clinical medicine (patient records)	310,314	1	37	2	2.57	yes	July 31, 2007
LOINC	[24]	Clinical observations and laboratory tests	46,406	1	3	3	2.85	no	Version 2.21 (no "natural language" names)
FMA	[25]	Human anatomical structures	~72,000	1	?	?	~1.50	yes	(not yet in the UMLS)
Gene Ontology	[28]	Functional annotation of gene products	22,546	1	24	1	2.15	yes	Jan. 2, 2007
RxNorm	[31]	Standard names for prescription drugs	93,426	1	2	1	1.10	no	Aug. 31, 2007
NCI Thesaurus	[34]	Cancer research, clinical care, public information	58,868	1	100	2	2.68	yes	2007_05E
ICD-10	[36]	Diseases and conditions (health statistics)	12,318	1	1	1	1.00	no	1998 (tabular)
MeSH	[38]	Biomedicine (descriptors for indexing the literature)	24,767	1	208	5	7.47	no	Aug. 27, 2007
UMLS Meta.	[41]	Terminology integration in the life sciences	1,4 M	1	339	2	3.77	n/a	2007AC (English only)

Bodenreider, O. (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Methods of Information In Medicine, 47, Supplement 1, 67-79.*

A. Holzinger 709.049 40/82 Med Informatics L03

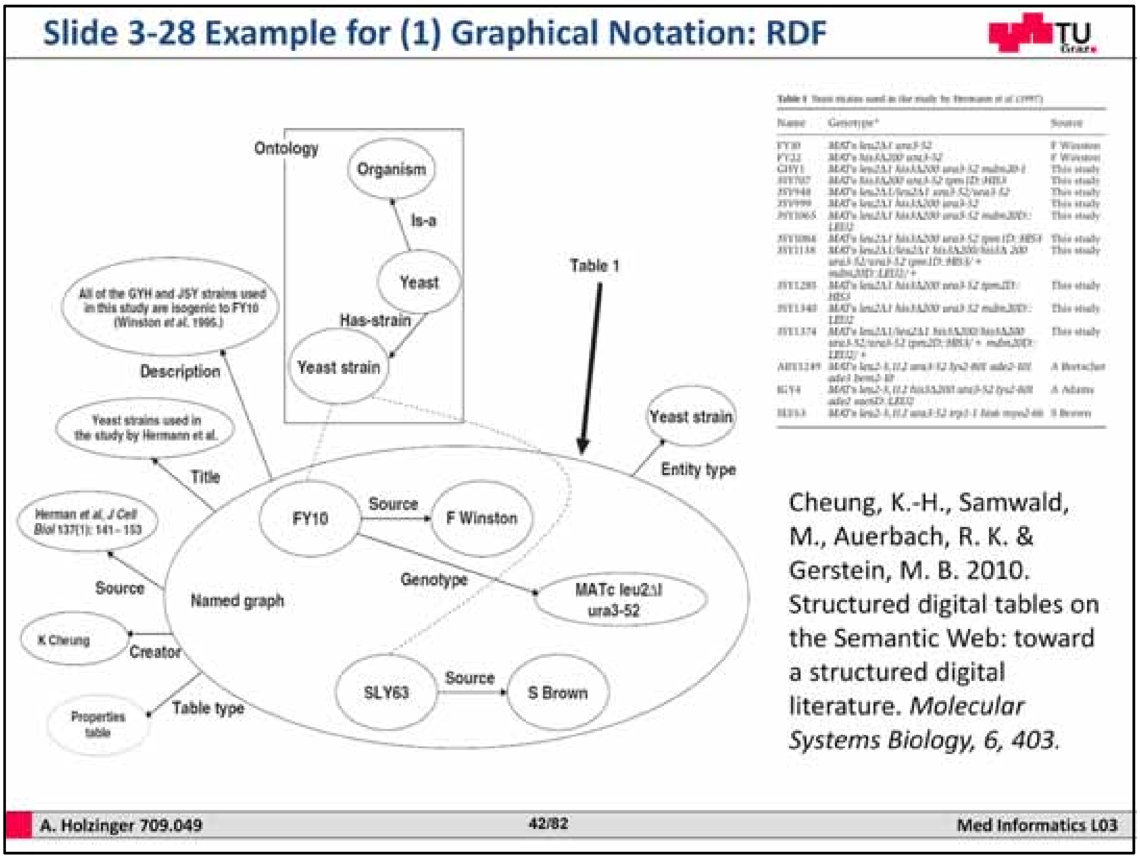
In this slide we see some biomedical ontologies, including scope, number of entities (concepts), distribution of the number of terms per entity (minimum, maximum, median and average), and existence of a sub-sumption hierarchy), based on information present in the UMLS version of 2007 (Bodenreider, 2008). Ontologies generally serve as a source of vocabulary, i.e., a list of names for the entities represented in these ontologies, however, collecting names is the function of terminology, not ontology, and ontology languages such as OWL, the Web Ontology Language, treat names as labels or annotations. In practice, however, most biomedical ontologies (with the notable exception of LOINC) provide lists of names for the entities they accommodate, in addition to properties and relations for these entities. The terminological component of biomedical ontologies is an important resource for natural language processing systems and supports knowledge management tasks such as annotation (or indexing) of resources, information retrieval, access to information and mapping across resources. However, the corpus of entity names present in biomedical ontologies covers only in part the lexicon of the domain (especially for languages other than English) and only forms the basis for managing term variation (Bodenreider, 2008).



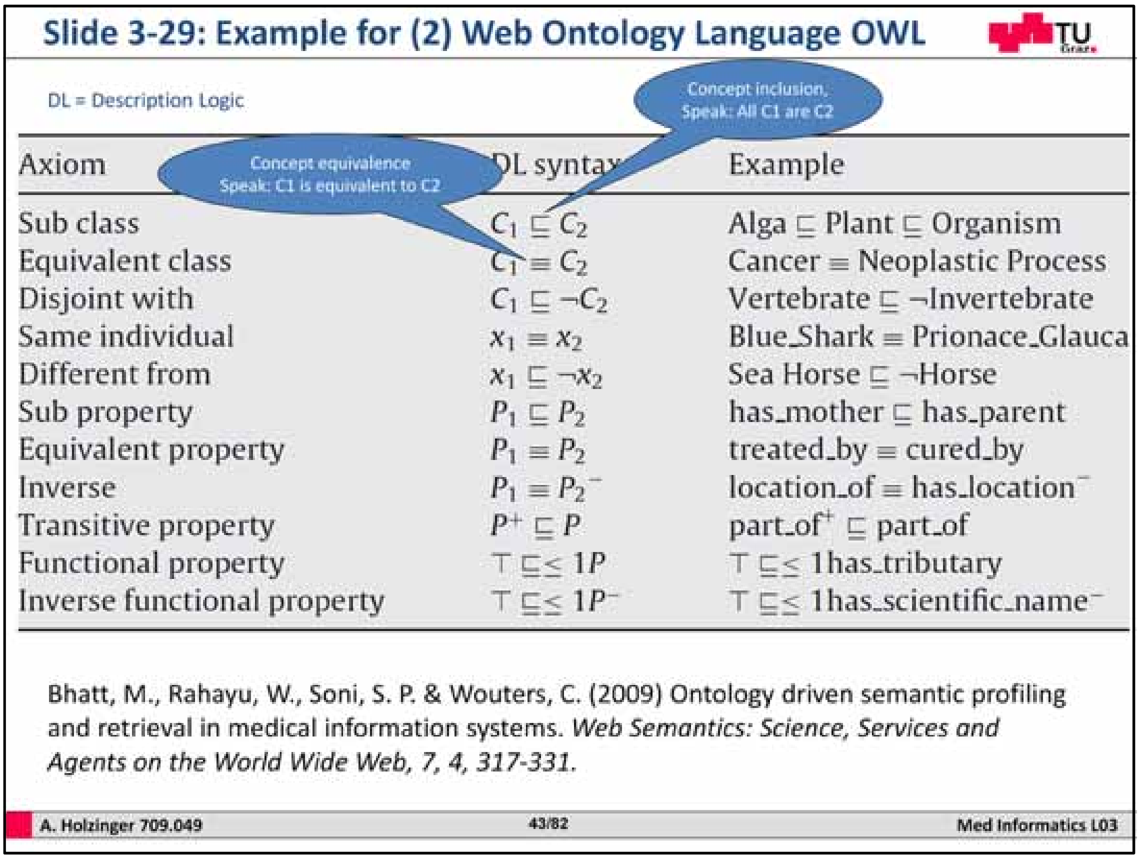
OIL = Ontology Interchange Language, DAML = DARPA Agent Markup Language, OWL = Web Ontology Language

Ontology languages are **formal languages** used to construct ontologies, allow the knowledge representation within specific domains and include reasoning rules, which support knowledge processing. Ontology languages are usually declarative languages, are almost always generalizations of frame languages, and are commonly based on either first-order logic or on description logic.
A coarse taxonomy is to determine between three concepts:
1) Graphical notations (semantic networks, topic maps, UML, RDF, …),
2) Logical based languages (e.g. description logics, OIL, DAML+OIL, OWL; rules, RuleML, LP/PROLOG; first order logics, KIF; conceptual graphs, syntactically higher order logics, Flogic, Non-Mon, modalities) and
3) Probabilistic/fuzzy approaches.
Note: KIF= Knowledge Interchange Format (e.g. Ontolingua), See:
http://www.ksl.stanford.edu/knowledge-sharing/kif;
http://www.isotopicmaps.org/sam/sam-model/

Fuzzy ontologies allow the modeling of real world environments using fuzzy sets mathematical environment and linguistic modeling. Therefore, fuzzy ontologies become really useful when the information that is worked with is imprecise.
Morente-Molinera, J. A., Pérez, I. J., Ureña, M. R. & Herrera-Viedma, E. 2015. Building and managing fuzzy ontologies with heterogeneous linguistic information. Knowledge-Based Systems, 88, 154-164.



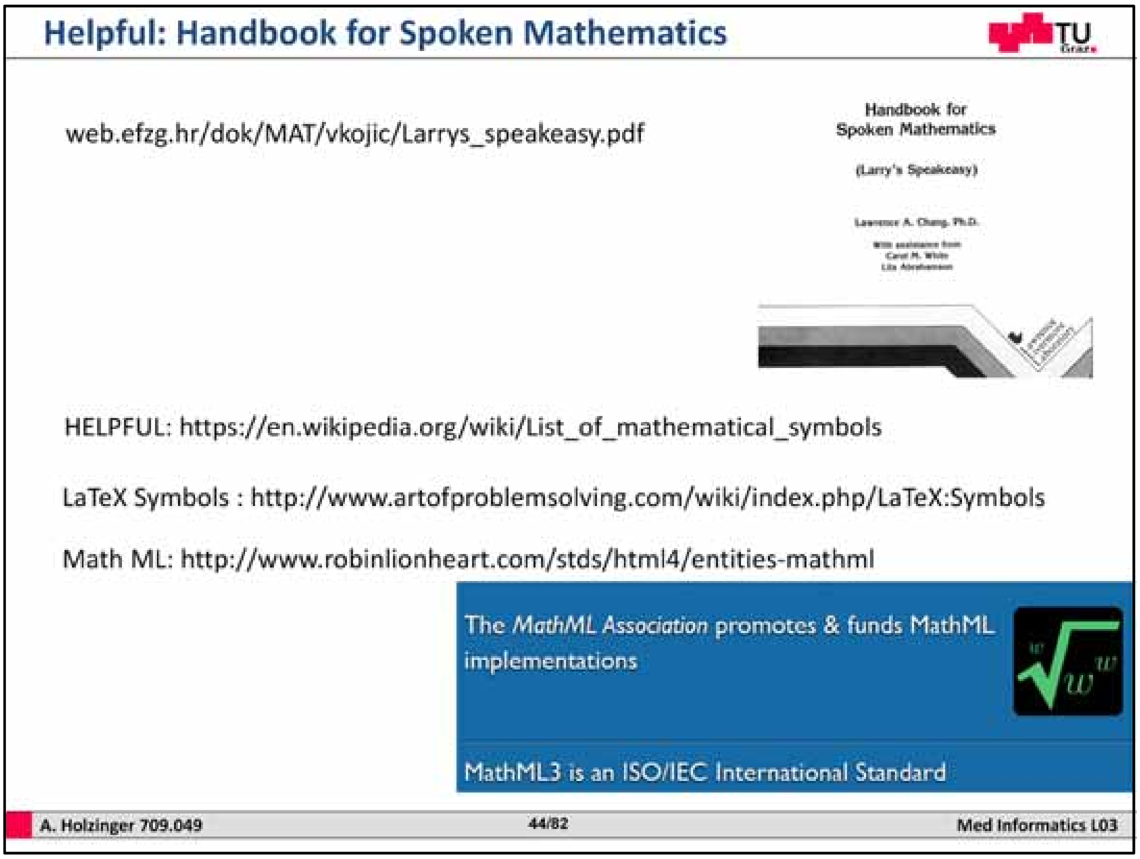
In this slide we can see the conversion of Table I into triples contained in a named graph (The source data for this figure is available at: www.nature.com/msb).
The table I is an example of a properties table (its canonical table counterpart has the same structure) and was obtained from a study to test whether the yeast gene, MDM20, is necessary for mitochondrial inheritance and organization of the actin cytoskeleton (Hermann, King & Shaw, 1997). It lists the different yeast strains in three columns (name, genotype, and source). Each table row corresponds to a specific yeast strain. We can apply the following rules to convert this table into RDF triples:
1. Each row is mapped to a subject
2. Each column header is mapped to a property
3. Each column value (cell) is mapped to a property value
The figure in the left depicts the mapping process and some of the mapping results. For the subject of each triple, we may check to see if it is an instance of an existing ontology class (represented using OWL or RDFS). For example, each subject (e.g. 'FY10') derived from Table I is an instance of (represented by a dotted line) the class 'yeast strain' in some organism ontology. Although the column name can be used to name the property, we may want to map it to some standard property name, if available. The generated triples represent a RDF graph. To this end, we use the named graph technique to identify the RDF graph generated from the table and to store the provenance information including the title, description (e.g. the table caption), creator, source (e.g. the paper), and so on. The properties (e.g. title, description, creator and source) are derived from the Dublin Core metadata standard http:// dublincore.org (Cheung et al., 2010).



The Web Ontology Language (OWL) is the most widely used ontology language, was developed by the W3C and thus designed specifically for use on the semantic web; it exploits existing web standards (XML and RDF), adding the familiar ontological primitives of object and frame based systems, and the formal rigor of a very expressive description logic (DL) that emerges from research in the field of Artificial Intelligence.

As we can see in Slide 3-29 and 3-30 the OWL consists of a rich set of knowledge representation constructs that can be used to formally specify medical-domain knowledge, which in turn can be exploited by description logic reasoners for purposes of inferencing, i.e., deductively inferring new facts from knowledge that is explicitly available.

The knowledge base (KB) of a typical DL based system comprises of two components, the TBox and the ABox; The TBox introduces the terminology, i.e., the vocabulary of an application domain (e.g., 'Neoplastic Process is-a Biological Function'), while the ABox contains assertions about named individuals in terms of this vocabulary ('Cancer is-a-instance of a Neoplastic Process'). The logical basis of the language means that reasoning services can be provided in order to make OWL described resources more accessible to automated processes. Formally, OWL is similar to a very expressive DL, with the OWL ontology corresponding to a DL terminology (TBox) whereas instance data pertaining to the ontology making up the assertions (ABox), therefore it is widely used in the medical domain (Bhatt et al., 2009).

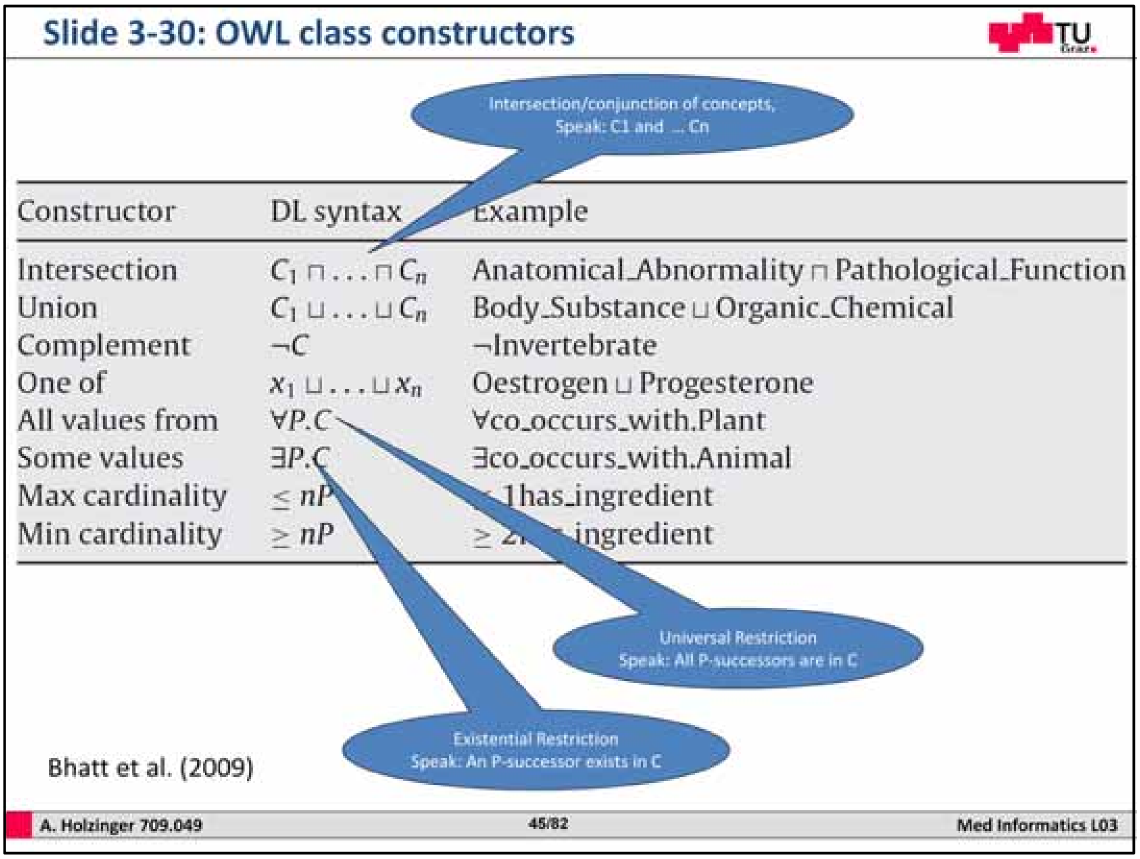


Mathematical Markup Language (MathML) is a mathematical markup language, an application of XML for describing mathematical notations and capturing both its structure and content. It aims at integrating mathematical formulae into World Wide Web pages and other documents. It is a recommendation of the W3C math working group and part of HTML5.

MathML is intended to facilitate the use and re-use of mathematical and scientific content on the Web, and for other applications such as computer algebra systems, print typesetting, and voice synthesis. MathML can be used to encode both the presentation of mathematical notation for high-quality visual display, and mathematical content, for applications where the semantics plays more of a key role such as scientific software or voice synthesis.

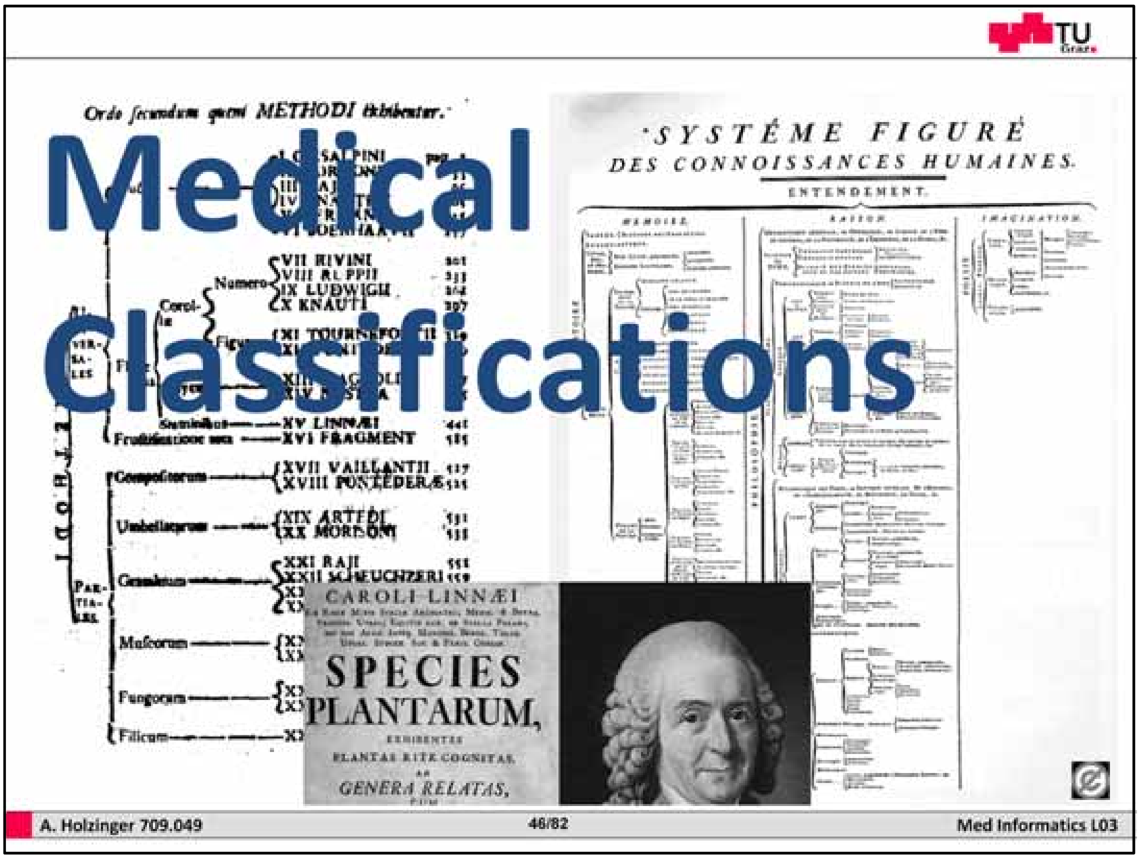
MathML is cast as an application of XML. As such, with adequate style sheet support, it will ultimately be possible for browsers to natively render mathematical expressions. For the immediate future, several vendors offer applets and plug-ins which can render MathML in place in a browser.

http://www.w3.org/Math/whatIsMathML.html



Constructor	DL syntax	Example
Intersection	$C_1 \sqcap \ldots \sqcap C_n$	Anatomical_Abnormality ⊓ Pathological_Function
Union	$C_1 \sqcup \ldots \sqcup C_n$	Body_Substance ⊔ Organic_Chemical
Complement	$\neg C$	¬Invertebrate
One of	$x_1 \sqcup \ldots \sqcup x_n$	Oestrogen ⊔ Progesterone
All values from	$\forall P.C$	∀co_occurs_with.Plant
Some values	$\exists P.C$	∃co_occurs_with.Animal
Max cardinality	$\leq nP$	≤ 1has_ingredient
Min cardinality	$\geq nP$	≥ 2has_ingredient

A Primer on OWL 2 as W3C recommendation from 11 December 2012 can be found here:
http://www.w3.org/TR/owl2-primer/



Classification is a general process in which ideas and objects are recognized, differentiated, and understood (semantics). A classification system is an approach to accomplishing classification.

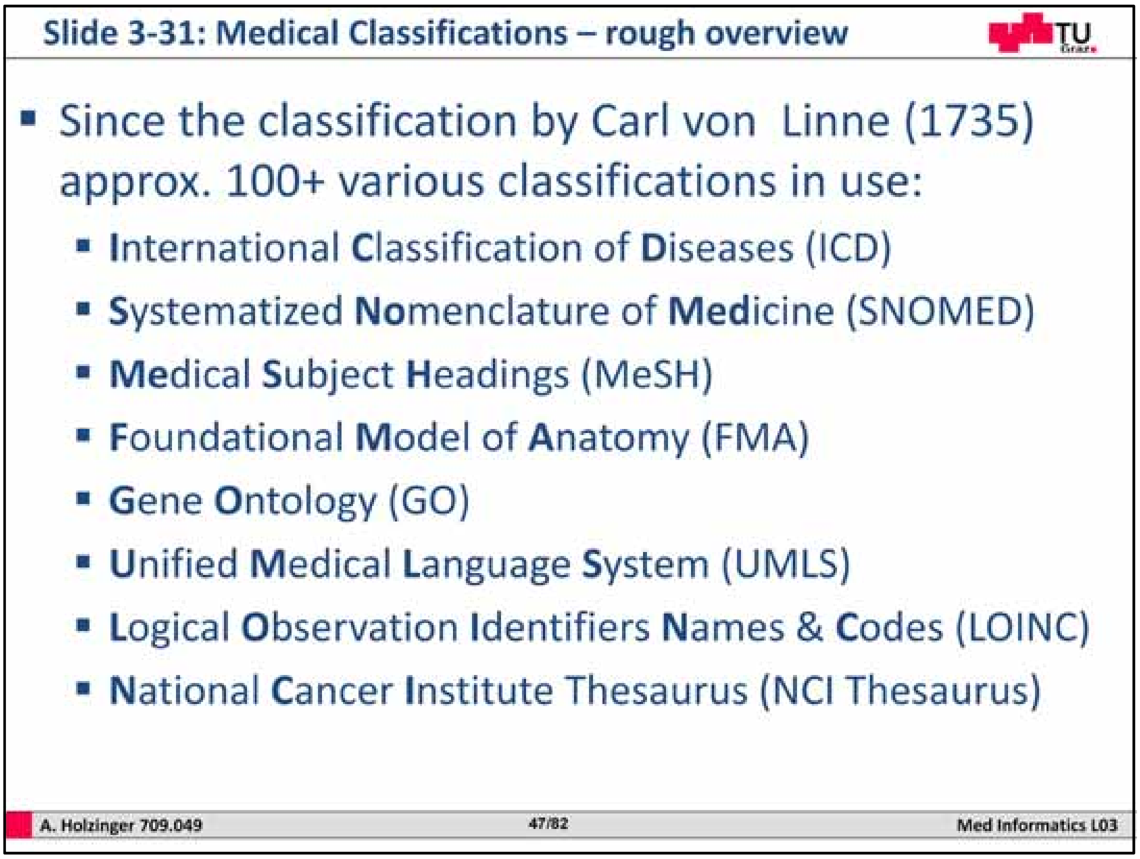
It goes back to Taxonomy, which is naming and classifying our surroundings to ensure a common understanding. E.g. Medicinal plant illustrations show up in Egyptian wall paintings from c. 1500 BC. The paintings clearly show that these societies valued and communicated the uses of different species, and therefore had a basic taxonomy in place.

Medical classifications are descriptions of medical diagnoses and procedures into universal medical codes.

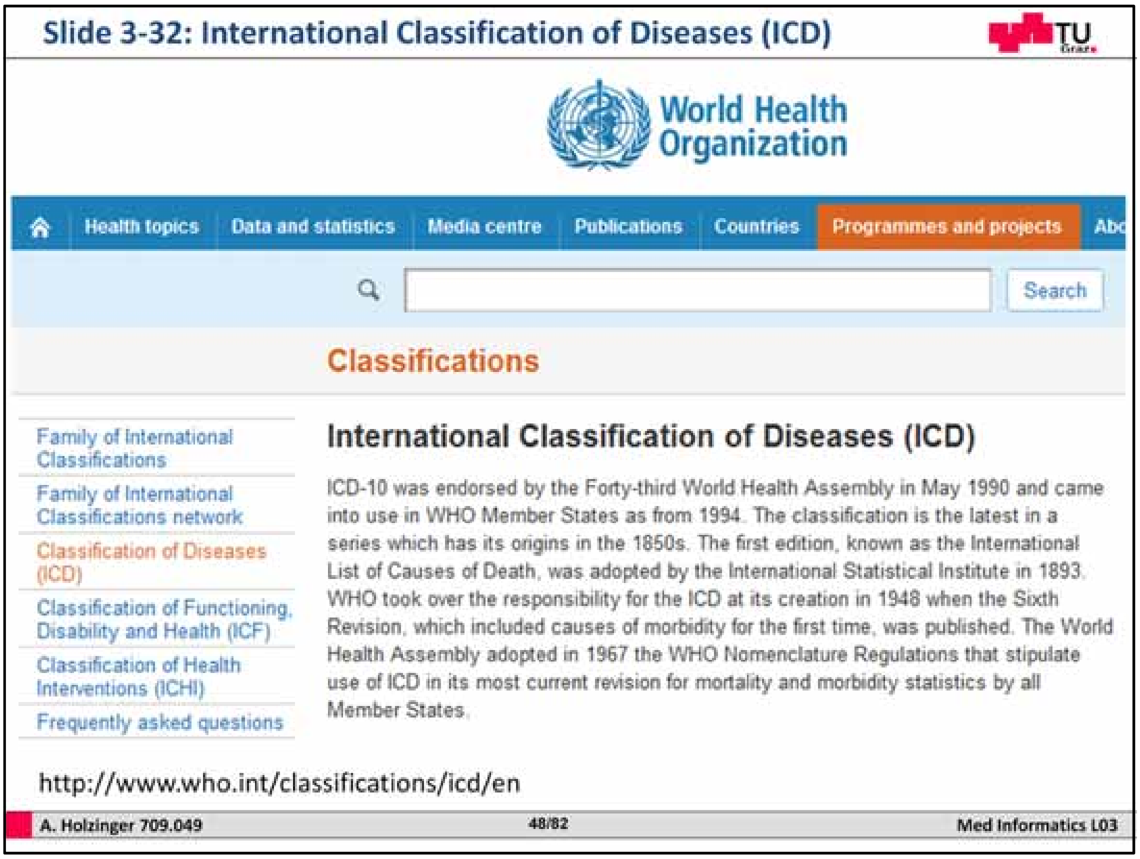
Nosology := from Ancient Greek νόσος (nosos), meaning "disease", and -λογία (-logia), meaning "study of-") deals with classification of diseases.

In the 18th century, the taxonomist Carolus Linnaeus, Francois Boissier de Sauvages, and psychiatrist Philippe Pinel developed an early classification of physical illnesses. Thomas Sydenham's work in the late 17th century might also be considered a nosology. In the 19th century, Emil Kraepelin and then Jacques Bertillon developed their own nosologies. Bertillon's work, classifying causes of death, was a precursor of the modern code system, the International Classification of Diseases.

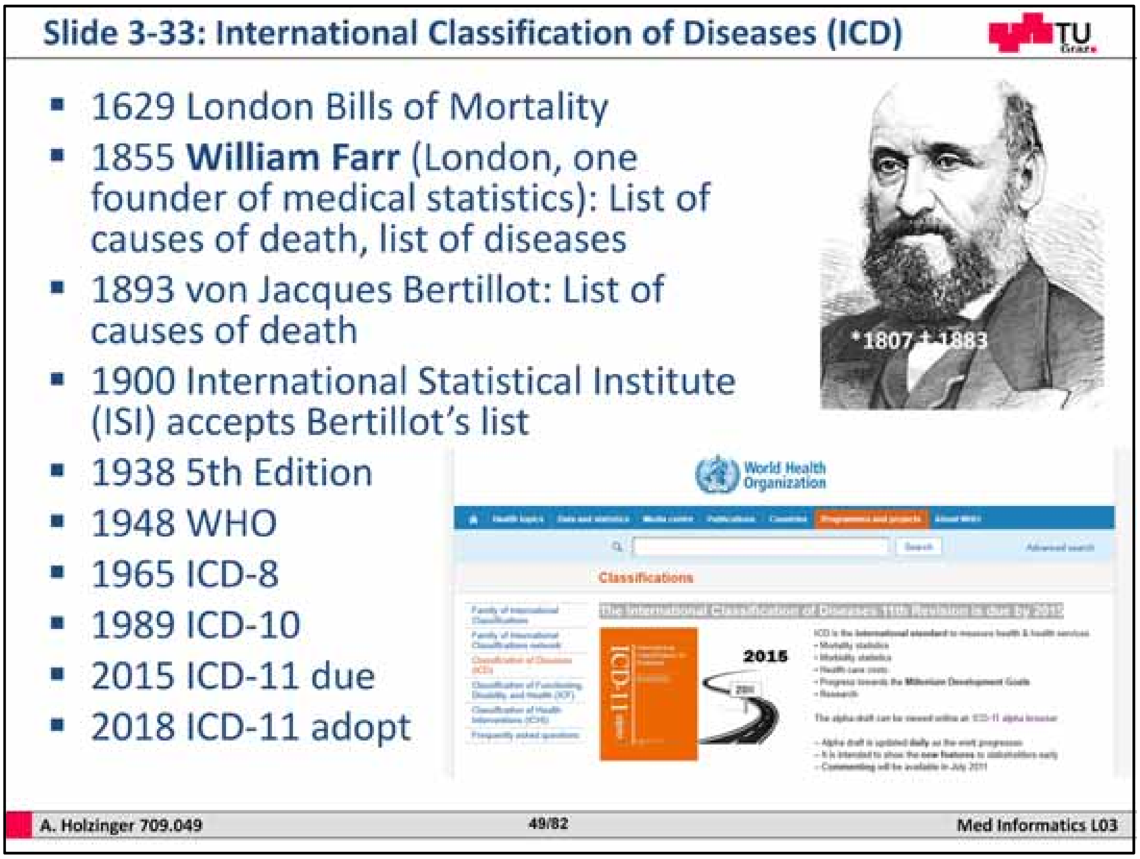
The early nosological efforts grouped diseases by their symptoms, whereas modern systems (e.g. SNOMED) focus on grouping diseases by the anatomy and etiology involved.



Medical classification, called coding by the professionals, is the process of transforming descriptions of medical diagnoses and procedures into a universal medical classification scheme.
A classification is a hierarchy of objects that conforms to the following principles (Berman, 2012):
1. The classes of the hierarchy have a set of properties that extend to every member of the class and to all of the subclasses of the class, to the exclusion of all other classes. A subclass is itself a type of class wherein the members have the defining class properties of the parent class plus some additional properties specific for the subclass.
2. In a hierarchical classification, each subclass may have no more than one parent class. The root class has no parent class.
3. The members of classes may be highly similar to each other, but their similarities result from their membership in the same class (i.e., conforming to class properties), and not the other way around (i.e., similarity alone cannot define class inclusion).
The father of classification was Carl von Linne (1707-1778) who began in 1735 with a classification of species. Today more than 100 various biomedical classifications are in use, for example:
International Statistical Classification of Diseases (ICD), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Medical Subject Headings (MeSH), Foundational Model of Anatomy (FMA), Gene Ontology (GO), Unified Medical Language (UMLS), Logical Observation Identifiers Names and Codes (LOINC), National Cancer Institute Thesaurus (NCI Thesaurus); Medical classification systems are used for a variety of applications in medicine, public health and medical informatics, including the reimbursement, e.g. based on diagnosis-related groups (DRG), but also for statistical analysis, therapeutic actions and knowledge engineering and decision support systems. Meanwhile, taxonomy is a science of classifying the elements of a knowledge domain,
and assigning names to the classes and the elements. In the case of terrestrial life forms, taxonomy involves assigning a name and a class to every species of life – on earth approx. 50 million species – a huge task. The central rules include (Berman, 2012):
1. All living organisms on earth contain DNA, which is transcribed into a less-stable, single-stranded molecule called RNA, which is translated into proteins. All living organisms replicate their DNA and produce more organisms of the same genotype.
2. All living organisms on earth can be divided into two broad classes:
the prokaryotes, the class that includes all bacteria; and eukaryotes.
3. The prokaryotes preceded the emergence of the eukaryotes, and the first
eukaryotes were built from the union of two or more prokaryotes.
4. Every eukaryotic organism that lives today is a descendant of a single eukaryotic ancestor.
5. Every organism belongs to a species that has a set of features that characterizes
every member of the species and that distinguishes the members of the species from organisms belonging to any other species.
For more information please refer to (Berman, 2012) and (Scamardella, 2010).

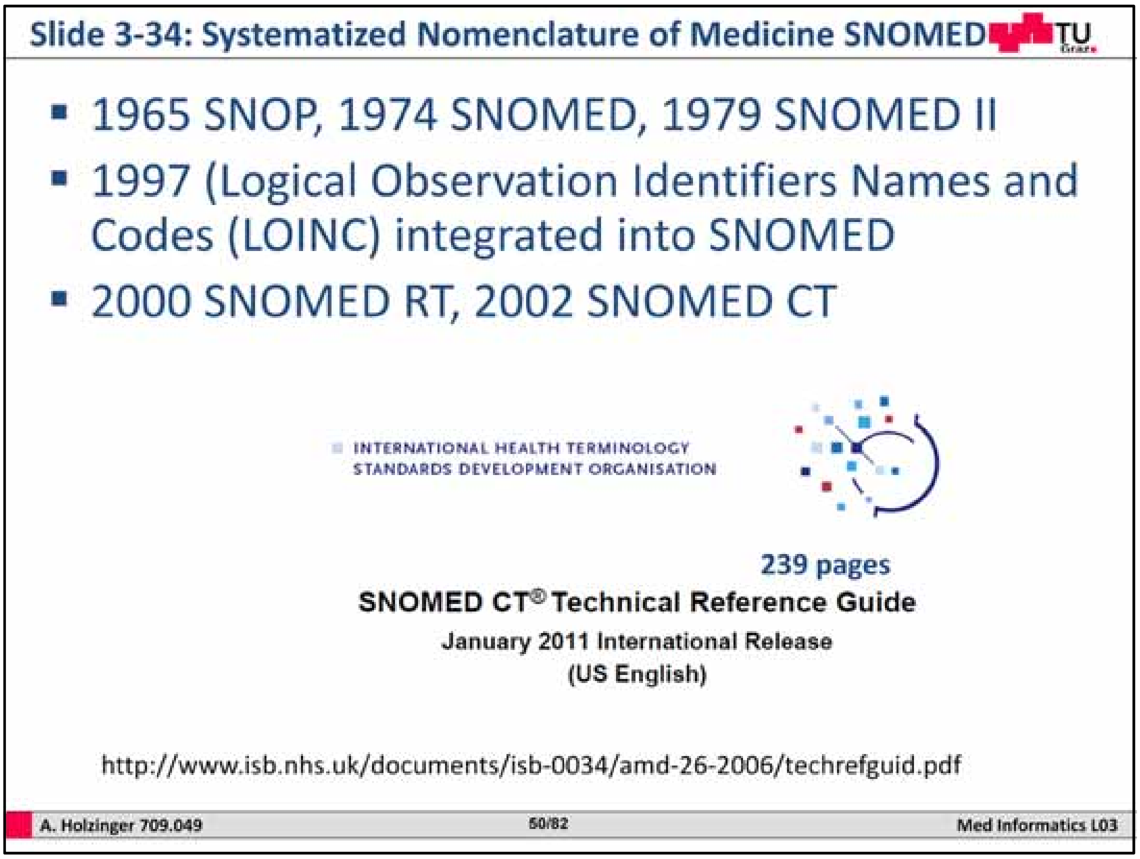


The International Classification of Diseases (ICD) is the standard diagnostic tool for epidemiology, health management and clinical purposes and includes the analysis of the general health situation of population groups. It is used to monitor the incidence and prevalence of diseases and other health problems as well as to classify diseases and other health problems recorded on many types of health and vital records including death certificates and health records. In addition to enabling the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes, these records also provide the basis for the compilation of national mortality and morbidity statistics by WHO Member States. It is used for reimbursement and resource allocation decision-making by countries. ICD-10 was endorsed by the Forty-third World Health Assembly in May 1990 and came into use in WHO Member States as from 1994.
The 11th revision of the classification has already started and will continue until 2015, for more details see: http://www.who.int/classifications/icd/en/

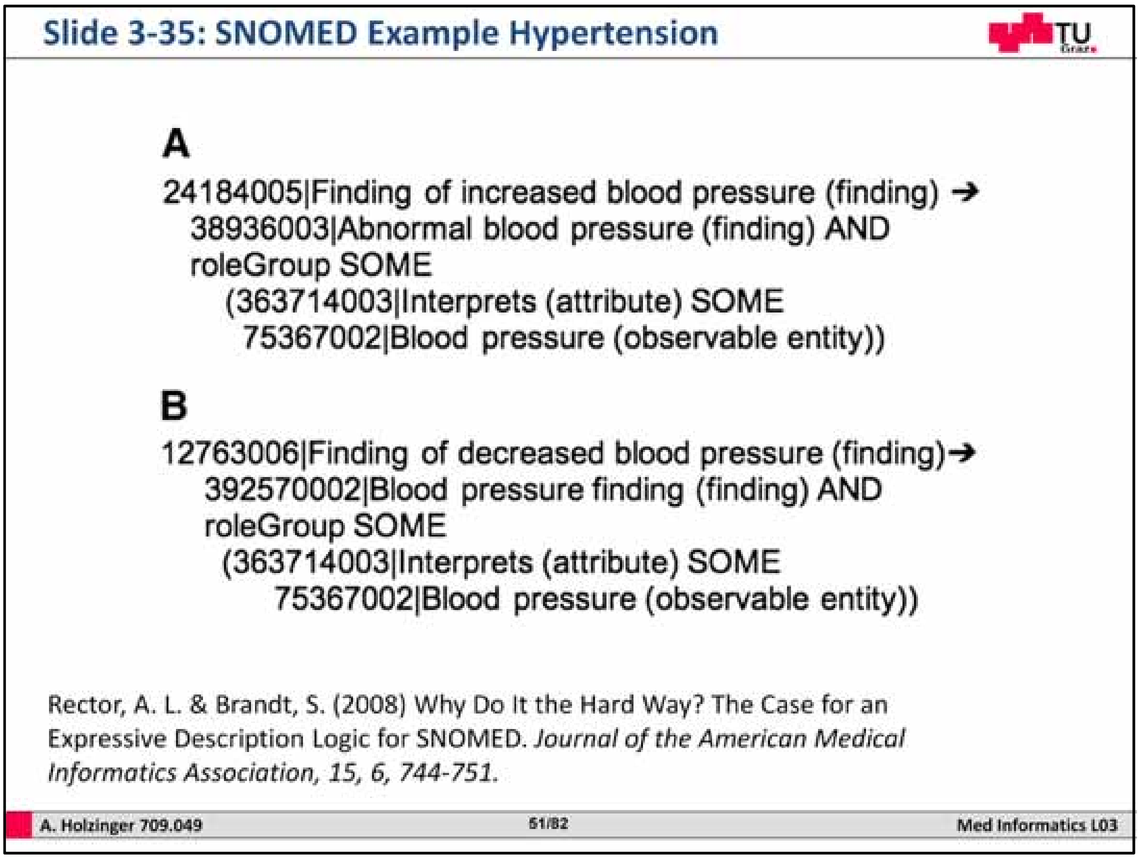


The oldest classification is the ICD, the roots can be traced back to:
1629 London Bills of Mortality
1855 William Farr (epidemiologist, London, one of the founders of medical statistics): List of causes of death, list of diseases
1893 Jacques Bertillot: List of causes of death
1900 International Statistical Institute (ISI) accepts the Bertillot list
1938 5th Edition
1948 WHO
1965 ICD-8
1989 ICD-10
2015 ICD-11 due
1965 SNOP, 1974 SNOMED, 1979 SNOMED II
1997 (Logical Observation Identifiers Names and Codes (LOINC) integrated into SNOMED
2000 SNOMED RT,
2002 SNOMED CT

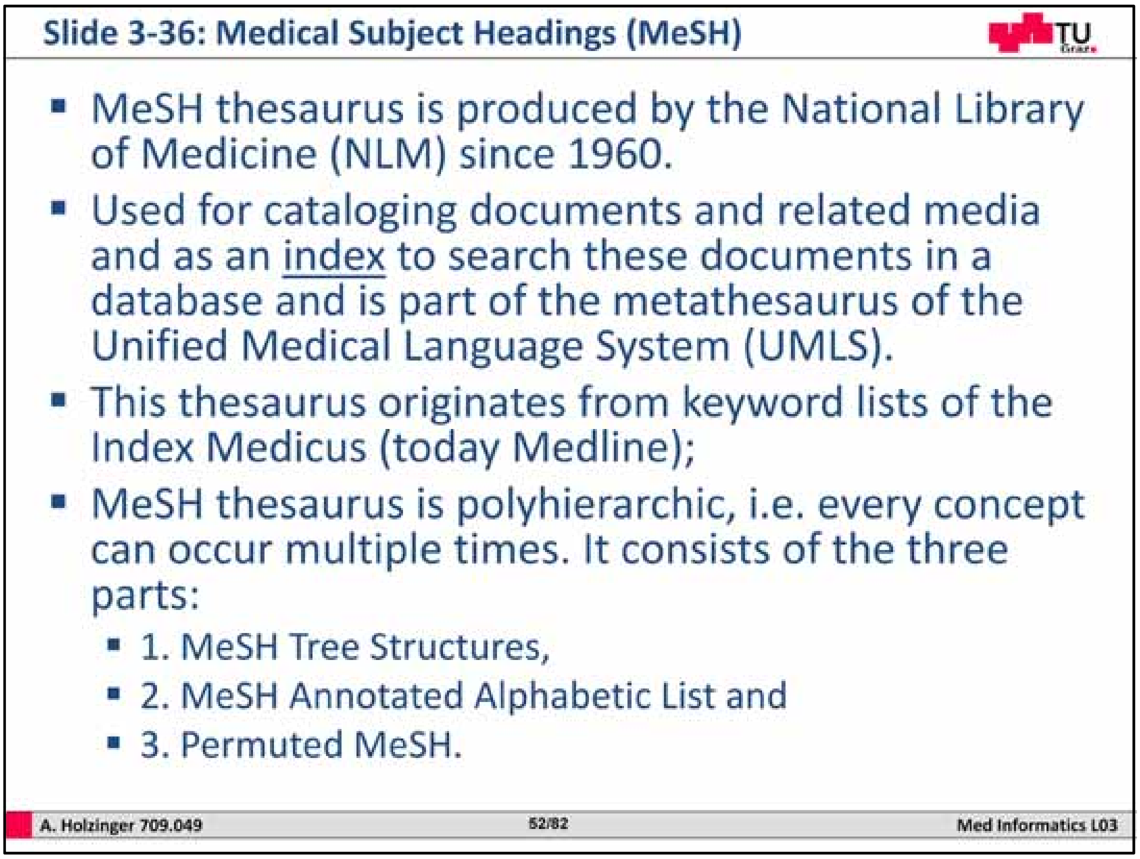
Jacques Bertillon, actually, introduced the Bertillon Classification of Causes of Death at a congress of the International Statistical Institute in Chicago in 1893 and thereof a number of countries and cities adopted his system, which was based on the principle of distinguishing between general diseases and those localized to a particular organ or anatomical site. Subsequent revisions represented a synthesis of English, German and Swiss classifications, expanding from the original 44 titles to 161 titles. In 1898, the American Public Health Association (APHA) recommended that the registrars of Canada, Mexico, and the United States also adopt it. The APHA also recommended revising the system every ten-years to ensure the system remained current with medical practice advances. As a result, the first international conference to revise the International Classification of Causes of Death took place in 1900.



SNOMED CT is the Systematized Nomenclature of Medicine Clinical Terms and covers diseases, clinical findings and procedures. Originally developed by the College of American Pathologists, the ownership of SNOMED CT was transferred to a new public body called the International Health Terminology Standards Development Organization (IHTSDO) in 2006. Presently, IHTSDO has 15 charter member countries with the common goal to develop, maintain and promote this terminology standard. The July 2009 version of SNOMED CT contains over 388,000 concepts, 1.14 million descriptions and 1.38 million relationships. There is a new release every six months through the National Release Centers of the respective charter member countries. With each release, there are changes that can affect the use of SNOMED CT within an organization's electronic patient record (EPR) systems. These include the fully specified name/preferred term, concept status, primitive/fully defined status, defining attributes, normal forms, and position within the "is a" hierarchy. Some of these changes may lead to unexpected consequences in subsequent encoding, equivalency and subsumption testing, and querying of a SNOMED CT (Lee, Lau & Quan, 2010).

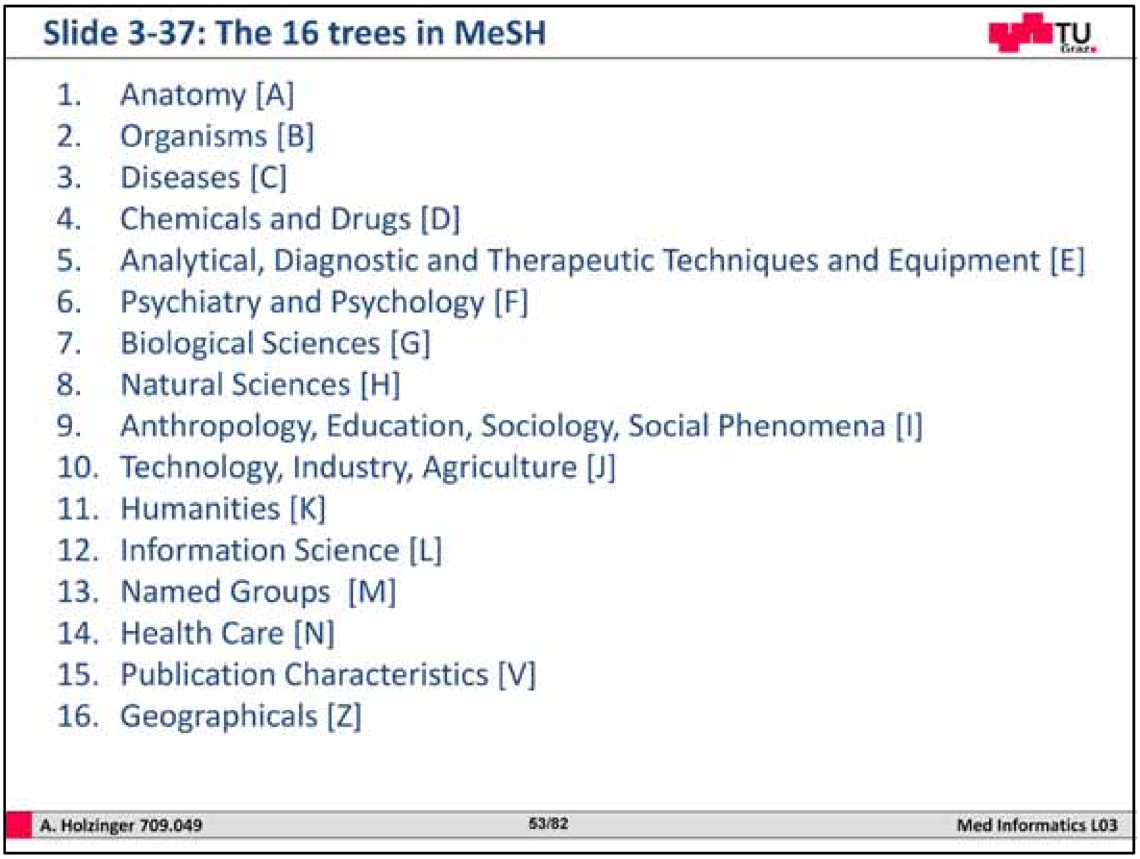


Here we see two examples: A. SNOMED Representation for increased blood pressure. B. SNOMED Representation for decreased blood pressure.
A big issue in clinical information systems is the distinction between observables and findings. Although there exists no universal consensus on the distinction, the term “observable” generally refers to an aspect of the patient that can be quantified or qualified, for example: blood pressure, skin color, body-mass index, etc.
A “finding,” on the other hand, usually refers to something which is either present or absent, possibly with additional qualification (diabetes, fractures, ...), or to the state of some observable such as “increased blood pressure” which likewise may be present or absent. In SNOMED, distinctions are made between the classes “finding” and “observable entity”. Figure A in Slide 3-35 makes this clear: the finding of increased blood pressure implies a finding of “abnormal blood pressure” that interprets the observable entity “blood pressure.” The fact that a finding of an “increased blood pressure” qualifies the blood pressure as abnormally high as opposed to abnormally low is not reflected at all in this expression! This is a common phenomenon. In many cases, most of the intended meaning behind concepts such as finding of increased blood pressure remains in the term name and is not reflected in a definition. This is even more obvious when comparing SNOMED’s (primitive) definition of a decreased blood pressure as shown in Figure B below (Rector & Brandt, 2008).



The MeSH thesaurus is produced by the National Library of Medicine (NLM) since 1960 and is used for cataloging documents and as an index to search these documents in a database, as part of the metathesaurus of the Unified Medical Language System (UMLS). This thesaurus originates from keyword lists of the Index Medicus (today Medline); MeSH is polyhierarchic, i.e. every concept can occur multiple times. It consists of the three parts:

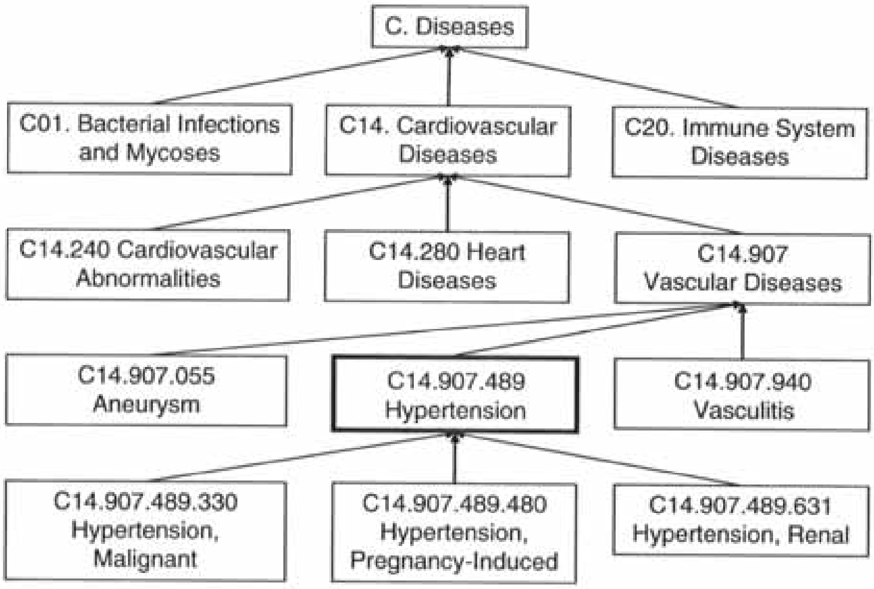
1. MeSH Tree Structures (see the Example in →Slide 3-37),
2. MeSH Annotated Alphabetic List and
3. Permuted MeSH.



The 16 trees in MeSH include:

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Natural Sciences [H]
9. Anthropology, Education, Sociology, Social Phenomena [I]
10. Technology, Industry, Agriculture [J]
11. Humanities [K]
12. Information Science [L]
13. Named Groups [M]
14. Health Care [N]
15. Publication Characteristics [V]
16. Geographicals [Z]

Slide 3-38: MeSH Hierarchy: e.g. heading Hypertension 1/2



Hersh, W. (2010) *Information Retrieval: A Health and Biomedical Perspective. New York, Springer.*

A. Holzinger 709.049 54/82 Med Informatics L03

This is an example for the MeSH Hierarchy for the heading Hypertension (Hersh, 2010) – the same example can be seen in the next slide as it looks originally in the Mesh Descriptor Database of the NLM.

Slide 3-39: MeSH Example Hypertension 2/2

National Library of Medicine - Medical Subject Headings

2011 MeSH

MeSH Descriptor Data

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

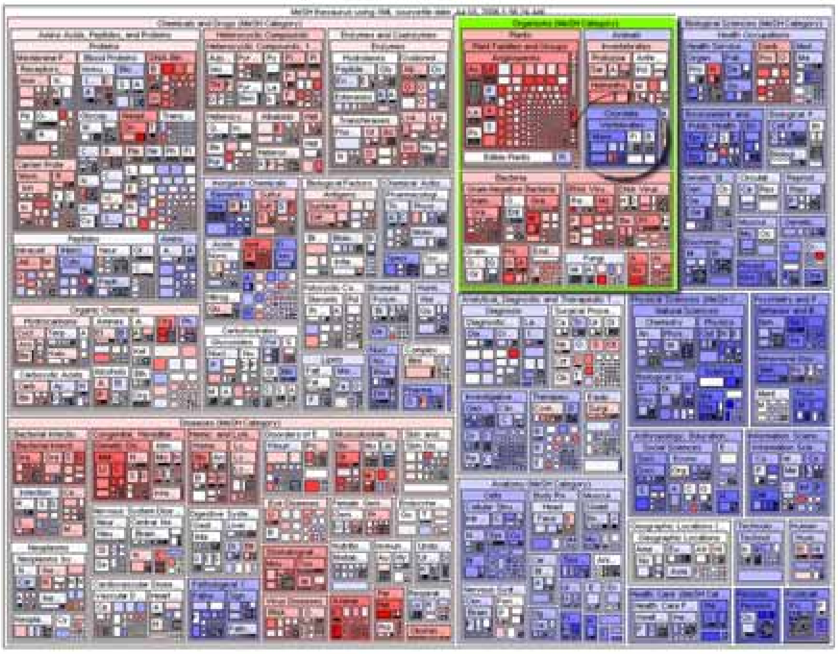
MeSH Heading	Hypertension
Tree Number	C14.907.489
Annotation	not for intracranial or intraocular pressure; relation to BLOOD PRESSURE: Manual 23.27; Goldblatt kidney is HYPERTENSION, GOLDBLATT see HYPERTENSION, RENOVASCULAR; hypertension with kidney disease is probably HYPERTENSION, RENAL, not HYPERTENSION; venous hypertension: index under VENOUS PRESSURE (IM) & do not coordinate with HYPERTENSION; PREHYPERTENSION is also available
Scope Note	Persistently high systemic arterial BLOOD PRESSURE. Based on multiple readings (BLOOD PRESSURE DETERMINATION), hypertension is currently defined as when SYSTOLIC PRESSURE is consistently greater than 140 mm Hg or when DIASTOLIC PRESSURE is consistently 90 mm Hg or more.
Entry Term	Blood Pressure, High
See Also	Antihypertensive Agents
See Also	Vascular Resistance
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Date of Entry	19990101
Unique ID	D006973

http://www.nlm.nih.gov/mesh/

A. Holzinger 709.049 55/82 Med Informatics L03

The example of Slide 3-38 as seen in the MeSH Database.
MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure, there are very broad headings such as "Anatomy". More specific headings are found at more narrow levels of the twelve-level hierarchy, such as "Ankle". In the 2013 version there are 26,853 descriptors and over 213,000 entry terms that assist in finding the most appropriate MeSH Heading, for example, "Vitamin C" is an entry term to "Ascorbic Acid." In addition to these headings, there are more than 214,000 headings called Supplementary Concept Records (formerly Supplementary Chemical Records) within a separate thesaurus. http://www.nlm.nih.gov/mesh

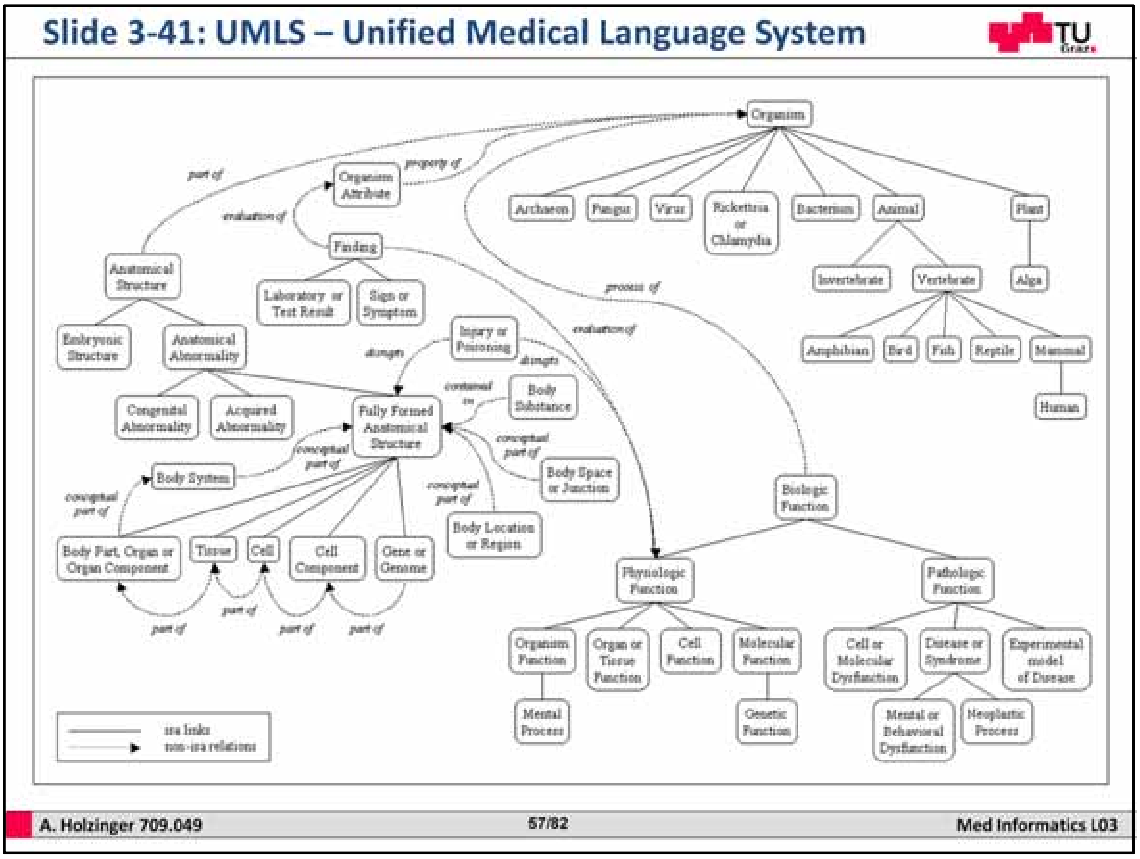
Slide 3-40: MeSH Interactive Tree-Map Visualization (see L 9)



Eckert, K. (2008) A methodology for supervised automatic document annotation. *Bulletin of IEEE Technical Committee on Digital Libraries TCDL, 4, 2.*

A. Holzinger 709.049 56/82 Med Informatics L03

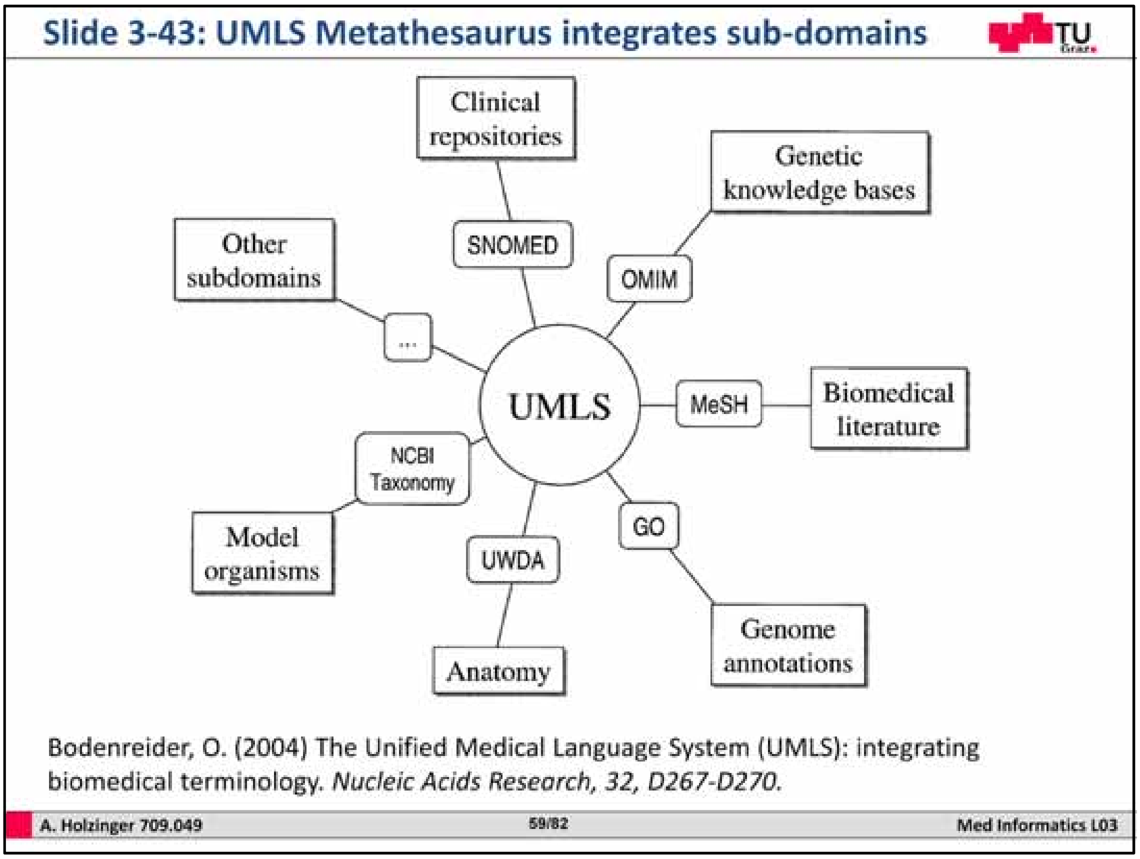
This is a very nice example of a possibility of visualization of such structures. We will discuss this in detail in →Lecture 9. The idea of such an approach is that the end-user has an idea of the overall structure (of the thesaurus) or selected parts of it. This example is a tree-map (Shneiderman, 1992): arbitrary trees are shown with a 2-d space-filling representation. With such a treemap, two additional aspects can be displayed beside the thesaurus structure: One is represented by the size of the partitions, the other by its colour. The hierarchy is visualized through the nesting of areas. The color of the different areas is used to represent the result of the different measures introduced above, for more details consult: http://www.ieee-tcdl.org/Bulletin/v4n2/eckert/eckert.html



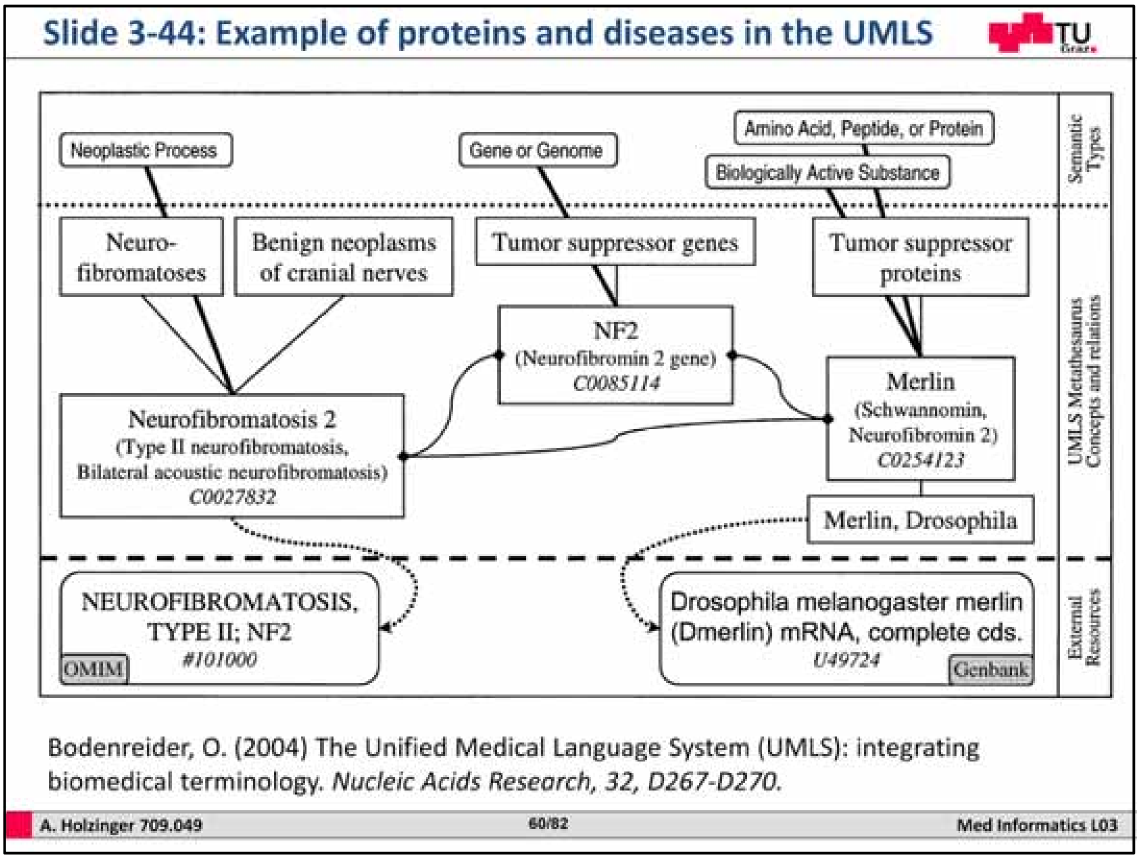
UMLS is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems (refer also to Slide 3-43). UMLS can be used to enhance or develop applications, such as electronic health records, classification tools, dictionaries and language translators.

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

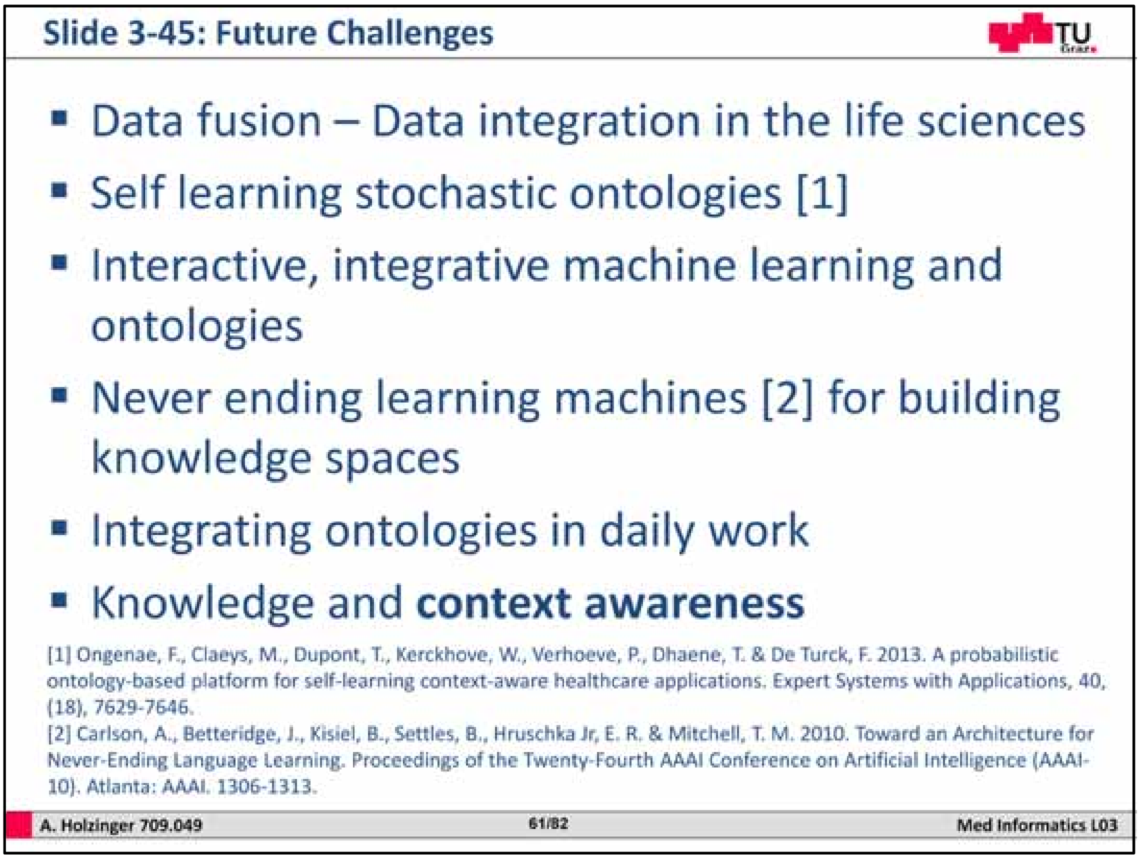
The Metathesaurus forms the base of the UMLS and comprises over 1 million biomedical concepts and 5 million concept names (!), all of which stem from the over 100 incorporated controlled vocabularies and classification systems. Some examples of the incorporated controlled vocabularies are ICD-10, MeSH, SNOMED CT, DSM-IV, LOINC, WHO Adverse Drug Reaction Terminology, UK Clinical Terms, RxNorm, Gene Ontology, and OMIM (to mention only a few).



In this slide we see the UMLS metathesaurus, integrating various other terminologies and serving as link between them and the subdomains they represent:
SNOMED - as link to clinical repositories;
OMIM -Online Mendelian Inheritance - as link to genetic knowledge bases;
MeSH - as link to biomedical literature (MEDLINE);
GO - as link used for the annotation of gene products across various model organisms;
UWDA University of Washington Digital Anatomist - as link to the Digital Anatomist Symbolic Knowledge Base;
NCBI - taxonomy used for identifying organisms;
Although the UMLS was not specifically developed for the needs of bioinformaticists, it includes terminologies used in bioinformatics. Integrated terminologies include the NCBI taxonomy, used for identifying organisms, and Gene Ontology, used for the annotation of gene products across various model organisms. The Metathesaurus also covers the biomedical literature with the MeSH, the controlled vocabulary used to index MEDLINE. Core subdomains such as anatomy, used across the spectrum of biomedical applications, are also represented in the Metathesaurus with the Digital Anatomist Symbolic Knowledge Base. Finally, the subdomain represented best is probably the clinical component of biomedicine, with general terminologies such as SNOMED International (and SNOMED-CT). Clinical genetics resources include the Online Mendelian Inheritance in OMIM represented in part, and the Online Multiple Congenital Anomaly/Mental Retardation (MCA/MR) Syndromes. Other categories of terminologies in the Metathesaurus include specialized disciplines (e.g. nursing, psychiatry) and components of the clinical information system (e.g. diseases, drugs, procedures, adverse effects). The figure illustrates how the UMLS Metathesaurus, by integrating these various terminologies, can serve as a link between not only the vocabularies, but also the subdomains they represent (Bodenreider, 2004).



For example, Neurofibromatosis 2 is an autosomal dominant disease characterized by tumors called schwannomas involving the acoustic nerve, as well as other features, where the disorder is caused by mutations of the NF2 gene resulting in the absence or inactivation of the protein product. The protein product of NF2 is commonly called merlin and functions as a tumor suppressor. Neurofibromatosis 2, NF2 and Merlin are concepts in the UMLS, for which the Metathesaurus provides many synonyms, including those listed above. In the slide we can see that these three concepts are linked by associative relationships: Each concept is part of a hierarchy of concepts. Neurofibromatosis 2 inherits from ancestors such as `Benign neoplasms of cranial nerves', which reflects the non-malignant behavior of schwannomas. Similarly, the function of NF2 is expressed through its direct parent `Tumor suppressor genes'. Semantic types from the UMLS semantic network provide a direct categorization to Metathesaurus concepts, making it easy to distinguish between the disease Neurofibromatosis 2 (Neoplastic Process) and the gene NF2 (Bodenreider, 2004).



A grand challenge is in data integration and data fusion in the life sciences and to make relevant data accessible to the clinical workplace. While there is much research on the integration of heterogeneous information systems, a shortcoming is in the integration of available data. Data fusion is the process of merging multiple records representing the same real-world object into a single, consistent, accurate, and useful representation (Bleiholder & Naumann, 2008), (McCray & Lee, 2013), (Horrocks, 2013).

Knowledge representation is an emerging field of artificial intelligence and stimulated ontologies in particular in the Web and its recent evolution, the so-called Semantic Web. The idea of the Semantic Web is consistent with some of the basic goals of knowledge representation. The vision is to enable semantic interoperability and machine interpretability of data sets from various sources and to provide the mechanisms that enable such data to be used to support the user in an automated and intelligent way. In order to establish a completely automated knowledge acquisition in the future, advances must be made both in the fields of natural language understanding and techniques of machine learning. The next generation of semantic applications will thus be characterized by the acquisition of knowledge from several sources instead of acquiring it from merely one source covering all the needs of target applications. Similar trends can also be expected in the use of knowledge available in existing ontologies. As it is not likely for a single ontology to satisfy all the needs of a certain application, the trends nowadays move towards ontology integration (also known as ontology alignment, matching or mapping). Integrating ontologies is one of the most complex and at the same time most important issues related to the practical implementation of Semantic Web. Consequently, the trend of integrating ontologies has lately gained substantial attention also in the research spheres and has actually become one of the most active fields of research. Although the results are very encouraging, so far integrated ontologies cannot be used in practice in most cases.

Due to the integration of knowledge from different sources, one of the challenges is ensuring a homogenous conceptualization of domains, as the contents of individual ontologies are very diverse and their vocabularies inhomogeneous, not to mention the differences in the quality of the presented knowledge.

Knowledge representation holds one of the key roles in the development of context awareness. Challenges in this field comprise of the formal presentation of the context, the determination of the formal relationships between different contexts of ontology use, the development of mechanisms for the selection of the appropriate context in a given situation and reasoning based on context. The development of reasoning based on context is especially important for user profiling, application personalization and mobility support. The examples of applications including the afore-mentioned areas are nowadays very popular social networks. To summarize, the results achieved in the domain of knowledge representation so far seem tentative and incomplete. Much work remains to be done. It is expected that under the auspices of Semantic Web and other accompanying concepts and visions, such as intelligent and personalized content retrieval, cloud computing, ubiquitous computing and, last but not least, artificial intelligence, the development of the field will continue (Jakus et al., 2013).

My DEDICATION is to make data valuable … Thank you!

Sample Questions (1)

- What is the proportion of structured/standardized versus weakly structured/non-standardized data?
- What are the benefits of standardized data?
- Which problems are involved in dealing with medical data?
- What is still a remaining big problem in the health domain … even with standardized data?
- What constitutes data standardization?
- What is the most used standardized data set in medical informatics today?
- Which are the three predominant ECG data formats?
- What is the advantage/disadvantage between binary data and XML data?
- What is the purpose of modeling biomedical knowledge?
- Provide examples for various abstraction levels of a Work Domain Model!
- What can be done with a Work Domain Model?
- What is the origin of ontologies?
- Please provide the classic definition of an ontology!
- What does domain semantics mean?
- What constitutes the classification of an ontology?

Sample Questions (2)

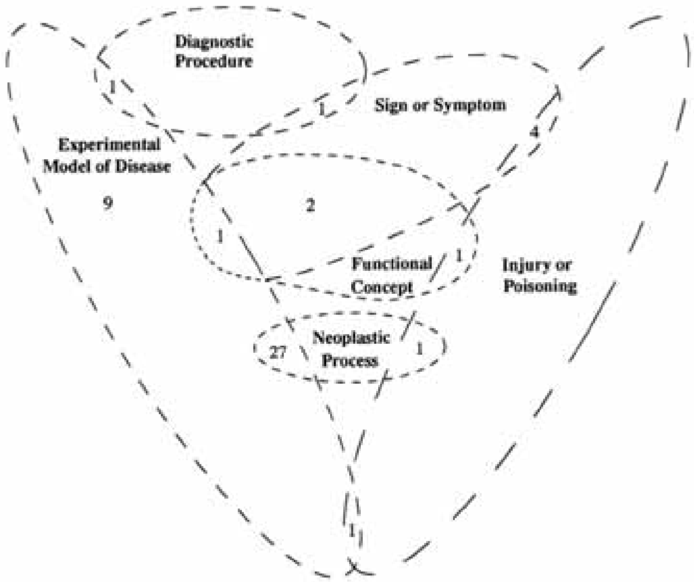
- Provide an overview about the most important biomedical ontologies!
- What are typical ontology languages?
- Please provide some examples of typical OWL axioms!
- What is a OWL class constructor?
- How do you start the development of an ontology?
- What are typical layers of abstraction – on the example of a Breast Cancer Imaging Ontology?
- What does "semantic enrichment" of a medical ontology mean?
- Within an ontology based architecture: what does the so called Knowledge Layer include?
- What are the roots of the ICD?
- What is the advantage of SNOMED-CT?
- What does polyhierachic thesaurus mean? Please provide an example for such a thesaurus!
- How can I expand queries with the MeSH Ontology?
- What is the major component of the UMLS?
- What is the main purpose of the Gene Ontology?

A. Holzinger 709.049 64/82 Med Informatics L03

Some useful links

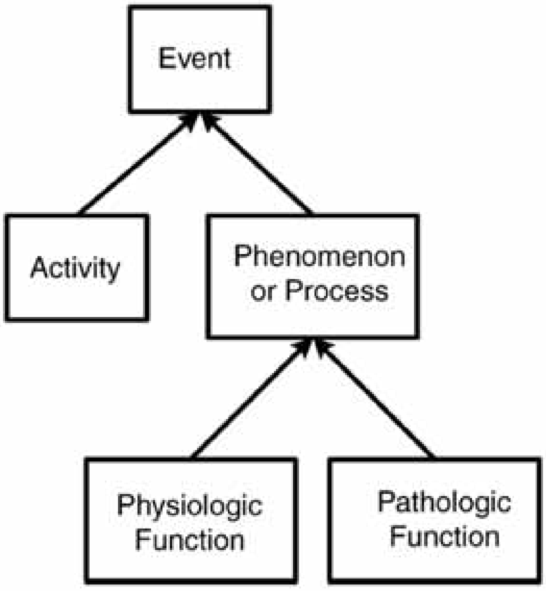
- http://wiki.hl7.org
- http://snomed.dataline.co.uk/
- https://github.com/drh-uth/MEDRank
- http://www.nlm.nih.gov/mesh/
- http://www.nlm.nih.gov/research/umls/
- http://www.geneontology.org/
- http://www.who.int/classifications/icd/en/

Backup-Slide: UMLS: Six semantic types and intersections



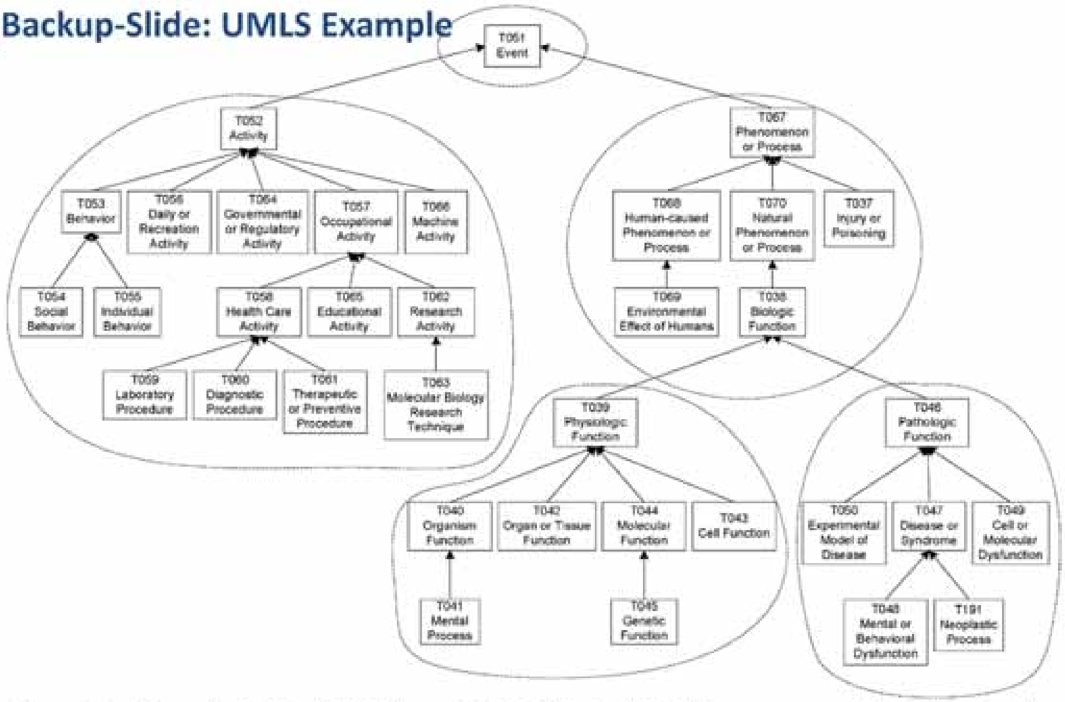
Gu, H., Perl, Y., Geller, J., Halper, M., Liu, L.-m. & Cimino, J. J. (2000) Representing the UMLS as an Object-oriented Database: Modeling Issues and Advantages. *Journal of the American Medical Informatics Association, 7, 1, 66-80.*

Backup-Slide: Metaschema hierarchy



Zhang, L., Hripcsak, G., Perl, Y., Halper, M. & Geller, J. (2005) An expert study evaluating the UMLS lexical metaschema. *Artificial Intelligence in Medicine, 34, 3, 219-233.*

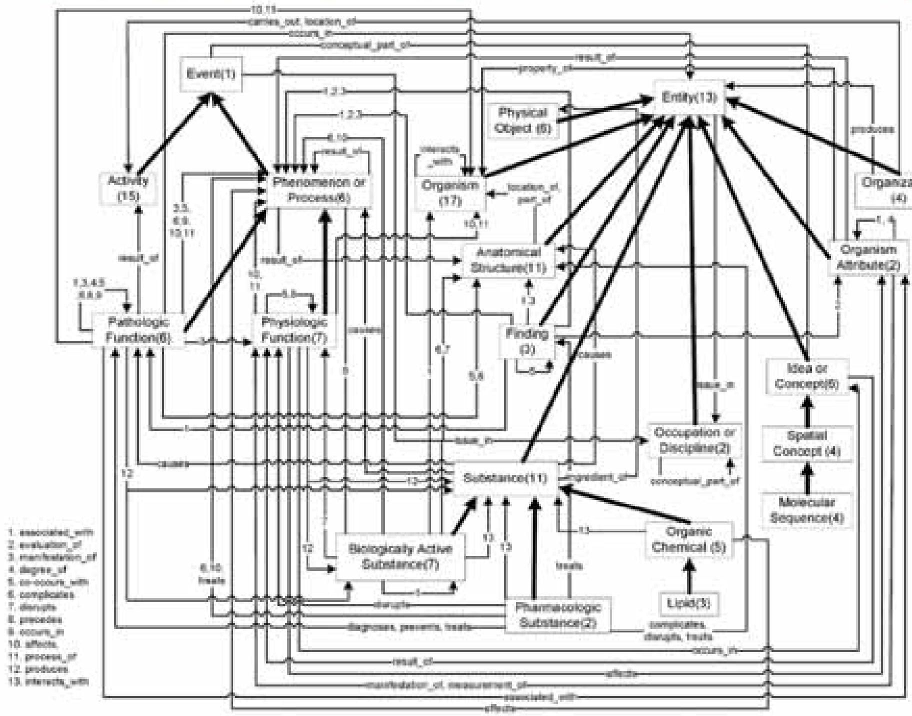
Backup-Slide: UMLS Example



Zhang, L., Hripcsak, G., Perl, Y., Halper, M. & Geller, J. (2005) An expert study evaluating the UMLS lexical metaschema. *Artificial Intelligence in Medicine, 34, 3, 219-233.*

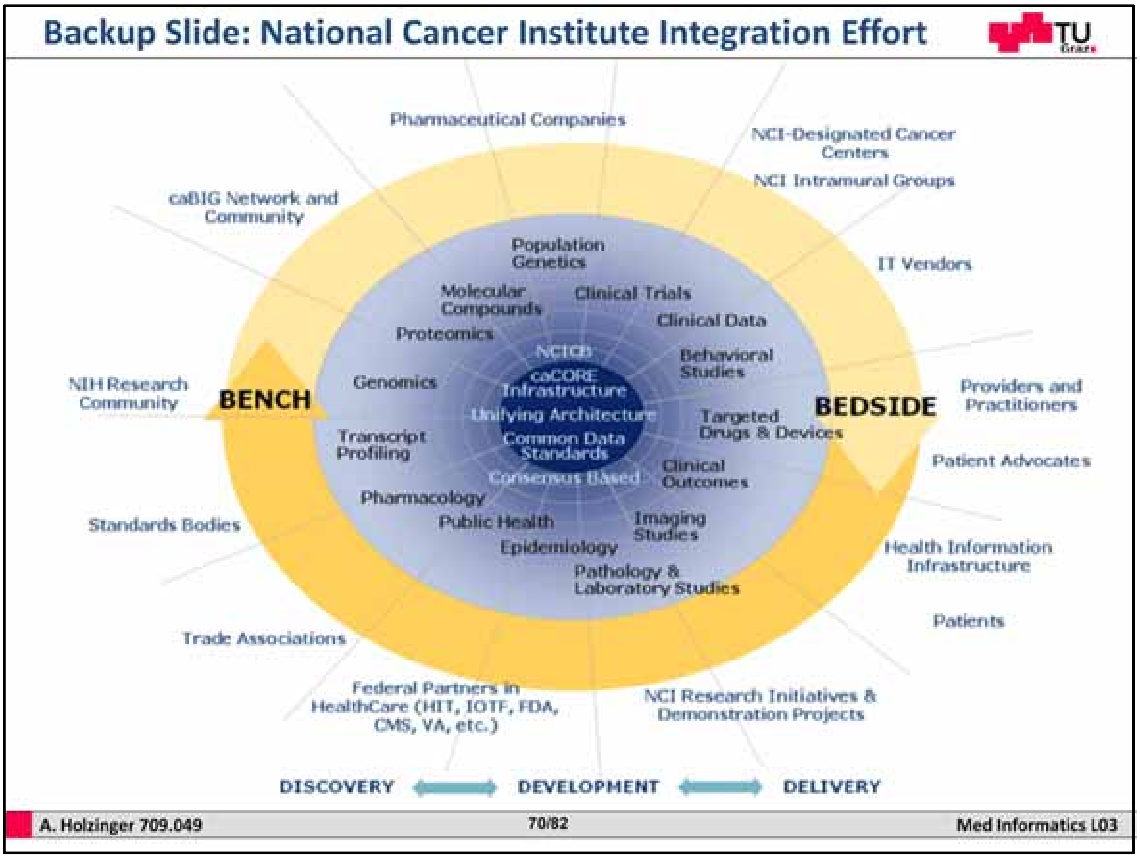
A. Holzinger 709.049 68/82 Med Informatics L03

Backup-Slide: Lexical Metaschema

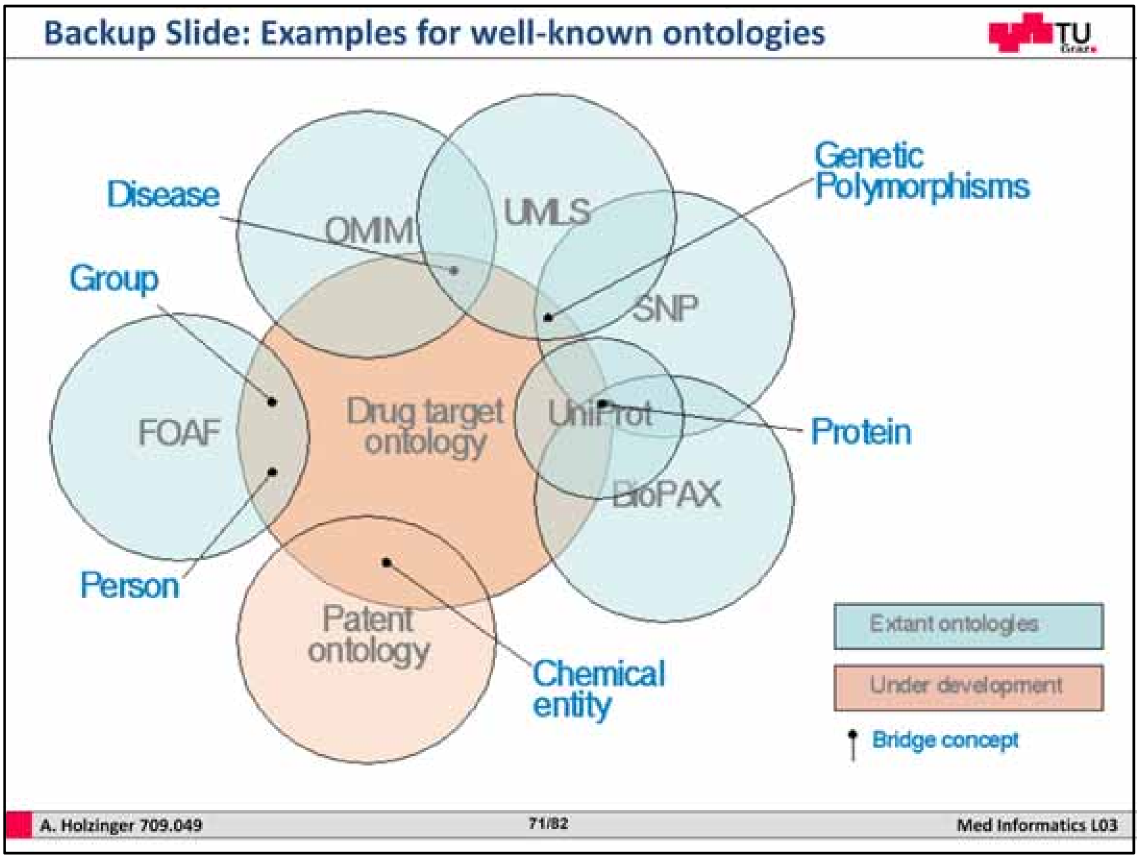


Zhang et al. (2005)

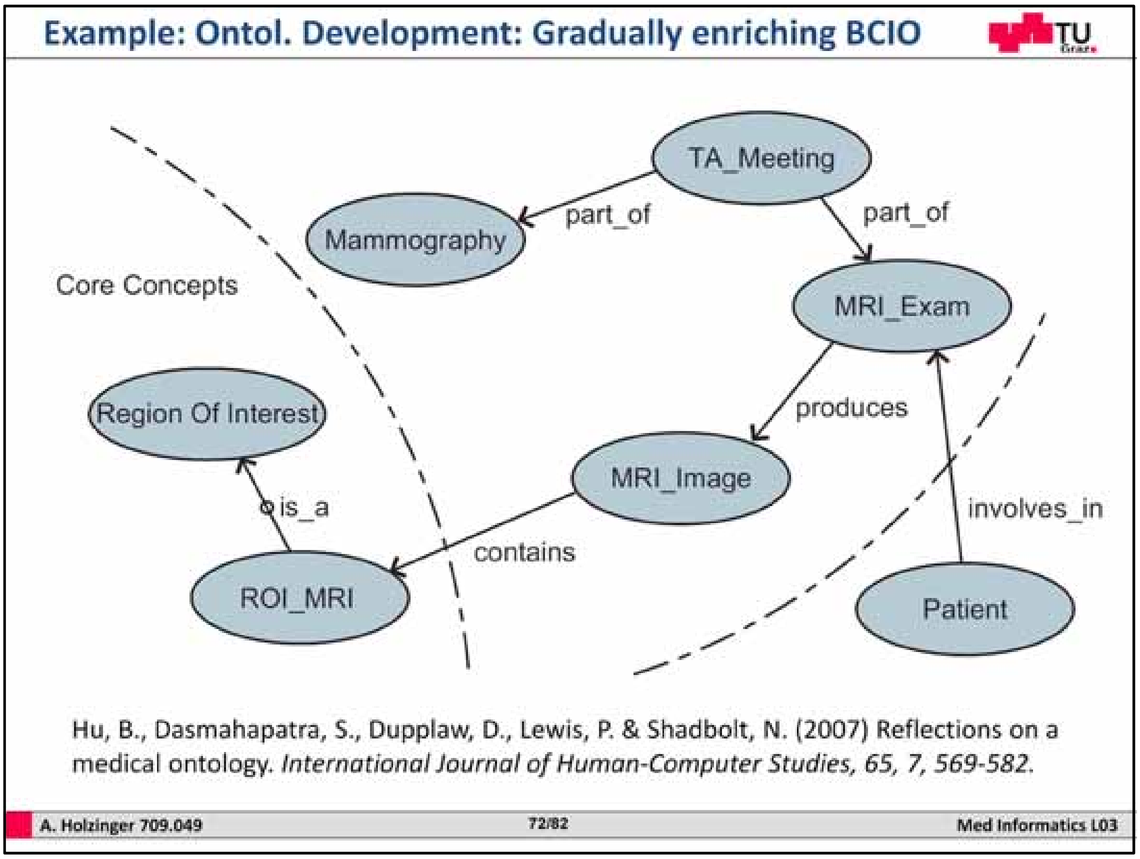
A. Holzinger 709.049 69/82 Med Informatics L03



http://ncicb.nci.nih.gov/about/initiatives

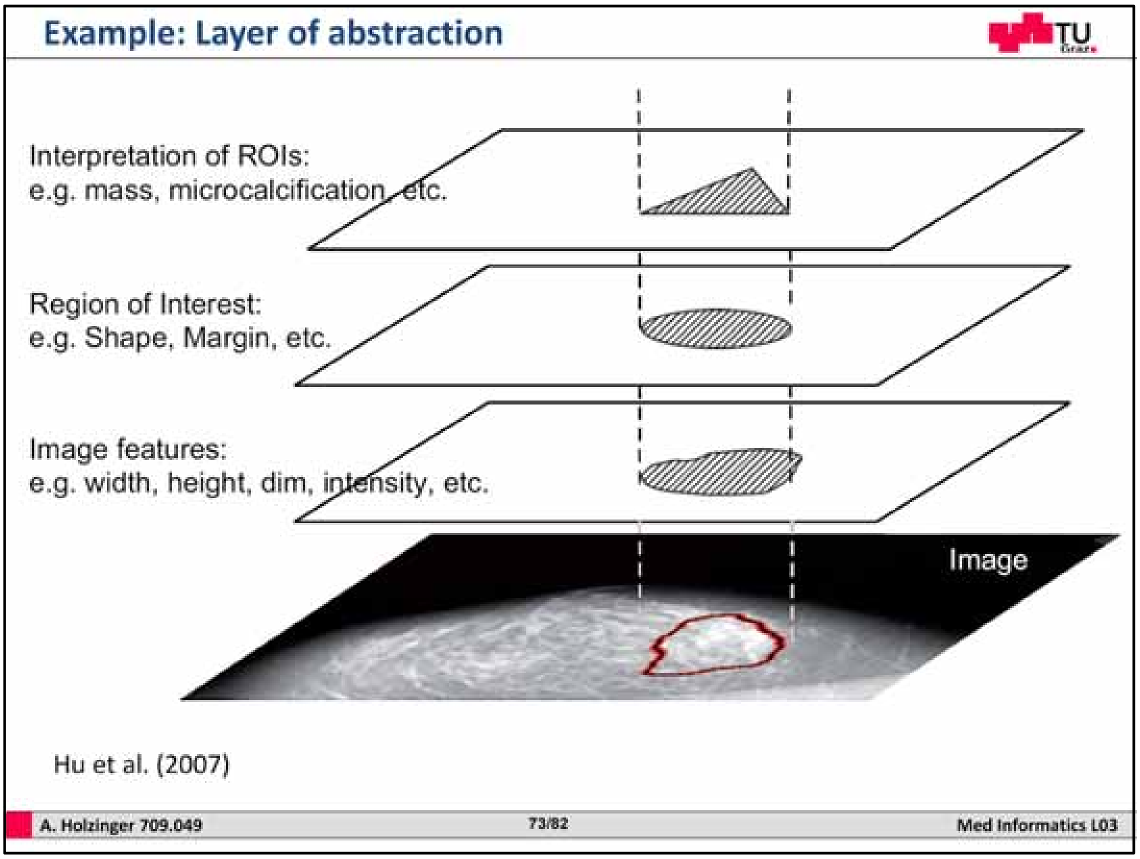


http://dig.csail.mit.edu/breadcrumbs/taxonomy/term/20



Breast Cancer Imaging Ontology (BCIO)

Ontology development could start with a limited and central set of entities from the domain of discourse, as gathered directly from domain experts or from the reports they write, and gradually enriching the initial set with knowledge whose relevance with those already included is over a predefined "threshold". Such a threshold could bedefined with respect to several criteria. One of the criteria we have chosen is linked to our mode of validation of the
ontology; namely, the possibility of capturing the key descriptive labels of cases that convey sufficient information to the specialist. The process of selecting and
deselecting relevant entities is itself supervised and reviewed by domain experts. We refer to the knowledge included in the initial set, i.e. those that are most central in the domain of discourse as target while the approach described above as target-driven. Breast cancer and breast cancer-related screening programs
have generated a large research literature. Close scrutiny of the literature reveals that during screening, attention converges on the abnormalities, which are
identifiable via the capabilities of different medical instruments and are provided as the evidence upon which conclusions are based. Furthermore, in the majority of screening protocols, patients are either flagged up during their routine X-ray screening or recommended by their family doctors with a follow-up X-ray examination. In both cases, abnormalities identified on X-ray images will be the starting point around which other evidence is gathered and accumulated to support a particular diagnostic decision. Such observations lead to treating identifiable ROIs on X-ray images, RegionOfInterest_Mammo in
BCIO. These are treated as the initial core components when constructing the ontology. It is thus an epistemologically located entry that ties the ontological description to its domain of interest, and is indispensable for any validation of the ontology. Other concepts are added either specifying other examinations as scheduled in the guideline to complete the screening protocol, e.g. MRI_Exam or to make more complete specifications of those concepts already included in the ontology, e.g. patient is added as the object of screening process, and a generalised category of medical examinations introduced. Such a process is illustrated in Fig. 1.
Hu, B., Dasmahapatra, S., Dupplaw, D., Lewis, P. & Shadbolt, N. 2007. Reflections on a medical ontology. International Journal of Human-Computer Studies, 65, (7), 569-582.



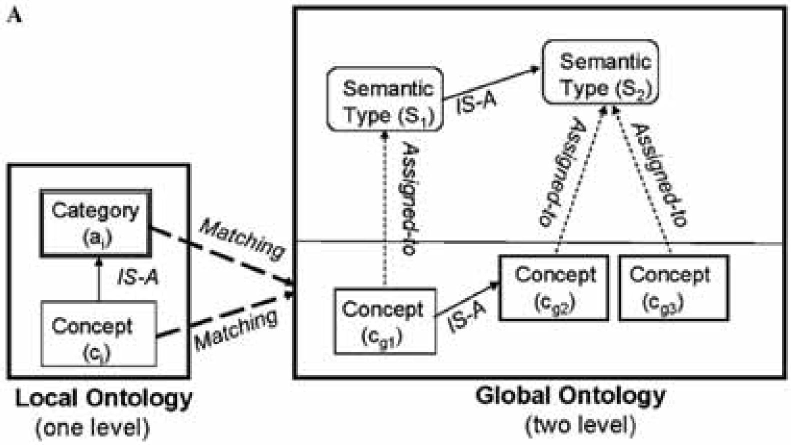
Researchers have been attacking this issue using different approaches: the multiplicativists consider colocalised entities as different individuals while the reductionists propose that they are different views of the same spatio-temporal entity. In order to explicitly model the correlation between abnormalities and ROIs, we present an eclectic mixture of the above extremes. In BCIO, we introduce several layers of abstraction (Fig. 2). Entities at each layer are abstracted from those at lower layers and the evidence for those at higher layers (see Section 4.1 for detailed discussion regarding multiple levels of abstraction).

Hierarchy of pathological concepts

An ontology is more than a simple classification of the domain of discourse; it is an aggregate of objects and processes as well as the connections among them. Hence, a ''full-fledged'' ontology (also referred to as heavyweight ontology) should demonstrate concepts, instances, conceptual hierarchies, and other relationships. However, in BCIO, we contend that because of ever expanding domain knowledge, which necessarily introduces a refocussing and elision of the totality of available descriptors historically attached to a condition, it is cumbersome to define a concept solely extensionally. Subjective knowledge, e.g. disease classification and prognostics, with attendant possibilities for intervention, needs to be included when objective observations are not sufficient to distinguish different concepts. For instance, although we can enumerate several symptoms of a particular breast disease, e.g. carcinoma in situ, it is impractical to list all known physical and pathological observations, not to mention those we have not discovered due to the limitation of current technologies and understanding. Such a situation is made even worse, if metastasis has occurred and other types of cancers and other anatomical loci are involved.

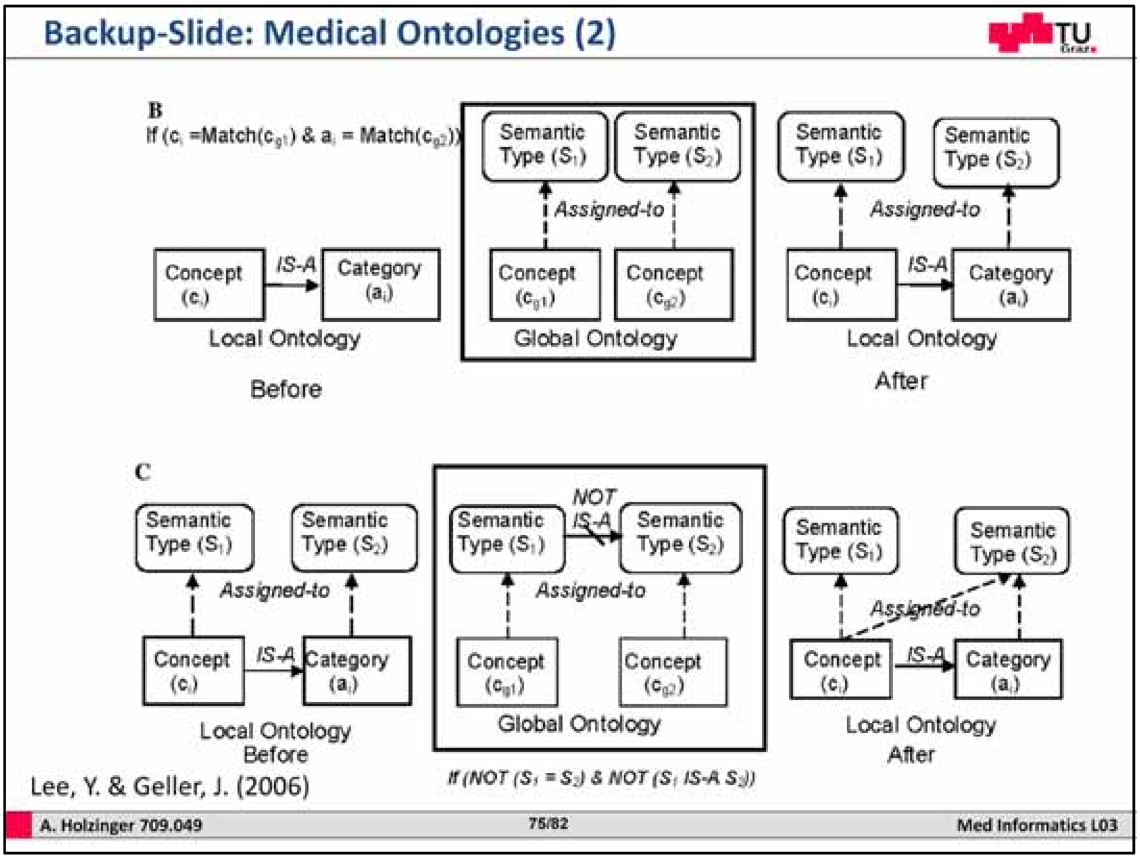
Hu, B., Dasmahapatra, S., Dupplaw, D., Lewis, P. & Shadbolt, N. 2007. Reflections on a medical ontology. International Journal of Human-Computer Studies, 65, (7), 569-582.

Backup-Slide: Medical Ontologies Semantic Enrichment

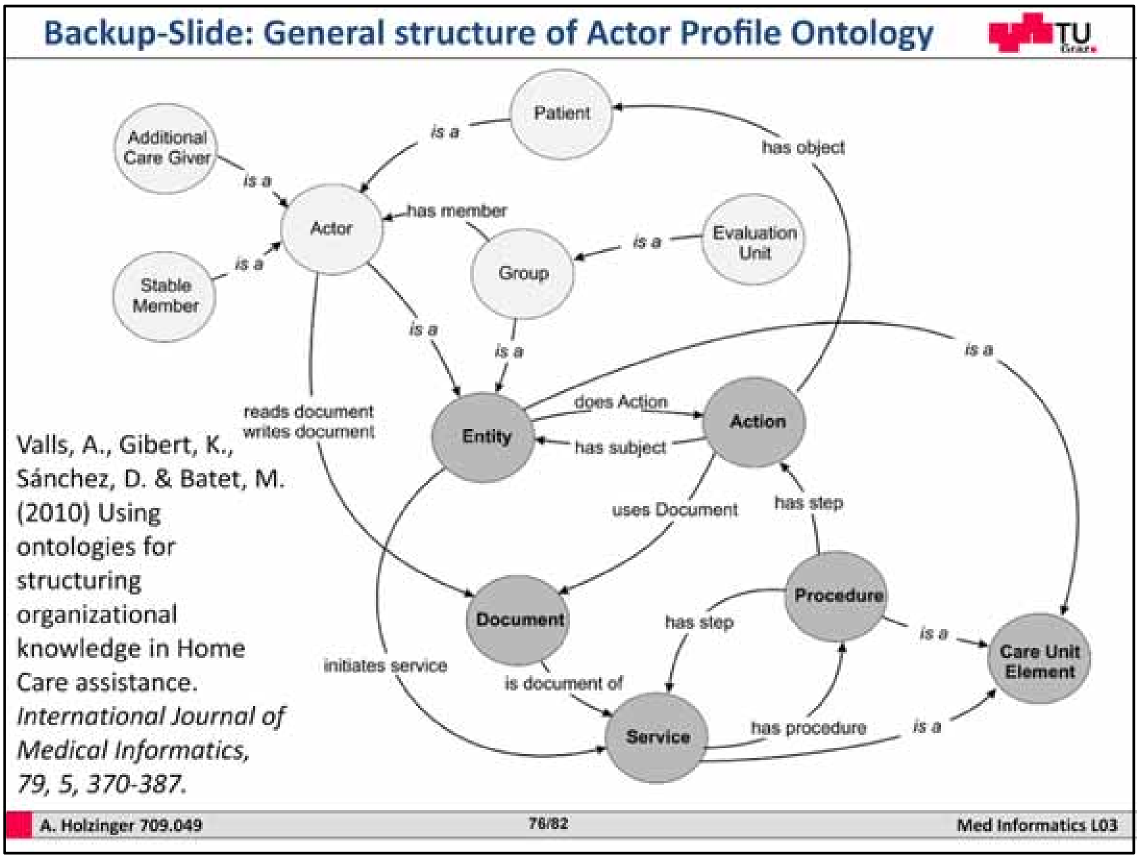


Lee, Y. & Geller, J. (2006) Semantic enrichment for medical ontologies. *Journal of Biomedical Informatics, 39,* ***2, 209-226.***

A. Holzinger 709.049 74/82 Med Informatics L03

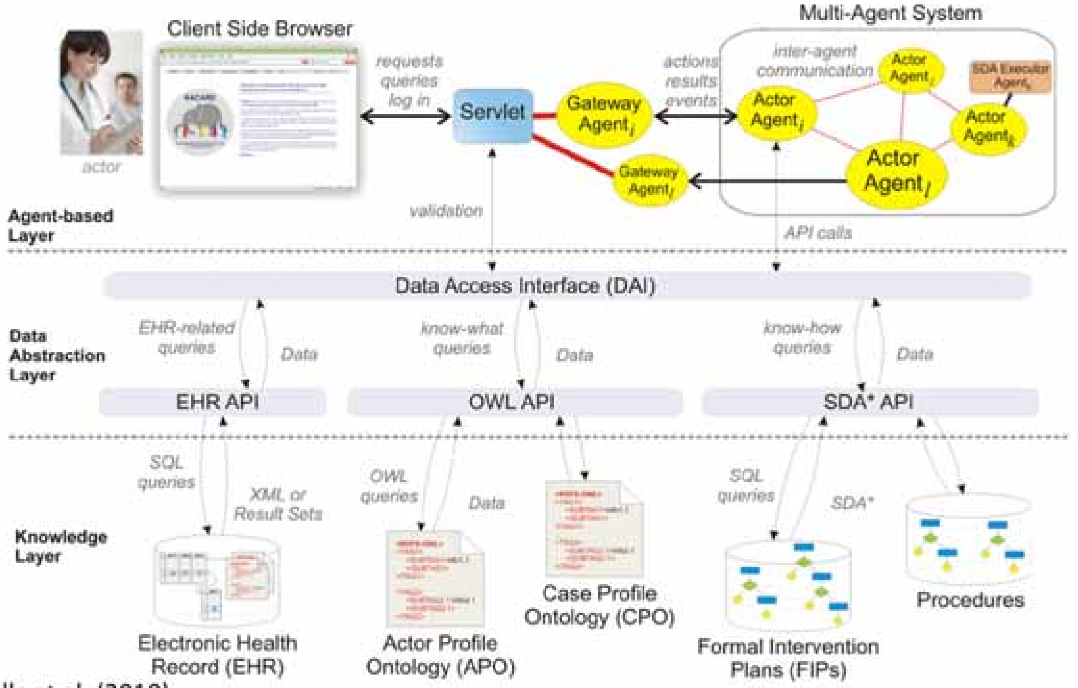


Step 1: concept matching. (B) Semantic assignment and (C) assignment propagation.
B shows the step of semantic assignment.



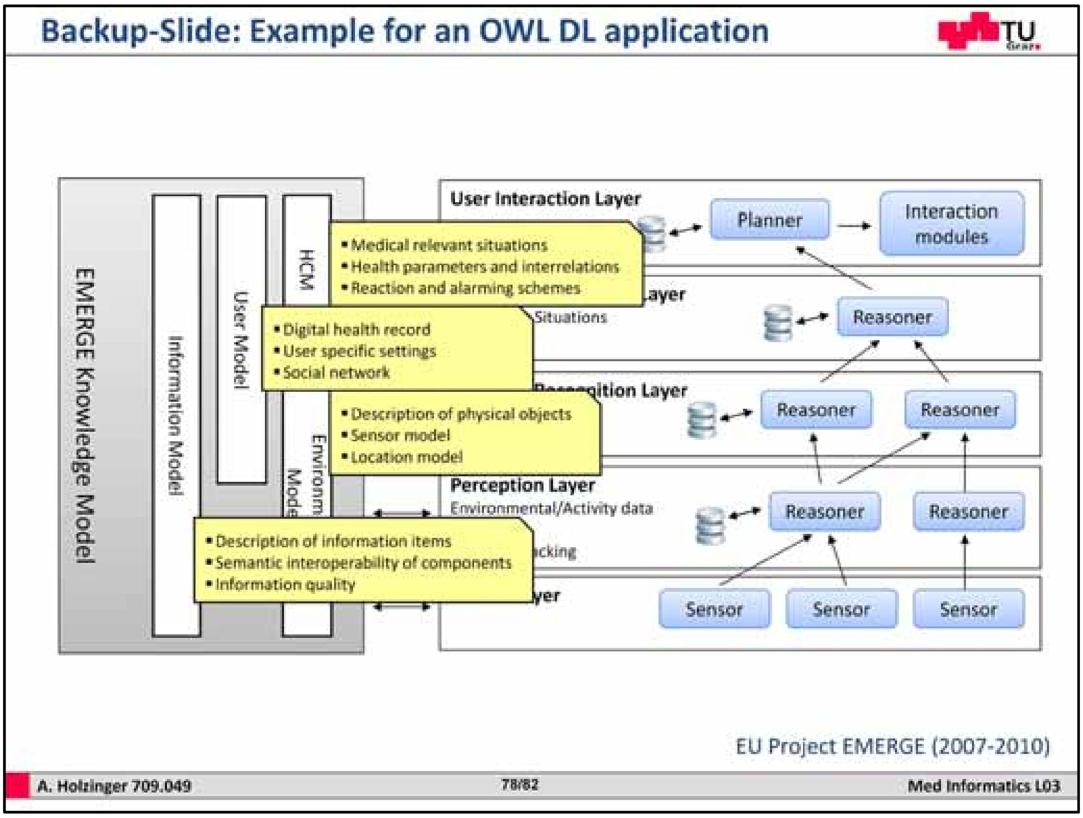
Formalization: According to the specification provided by medical experts, ontological entities were organized in several concept hierarchies (the top level elements of these
concept hierarchies are shown in dark grey in this slide),
together with the relationships between them.
The ontology obtained after this process can e.g. be coded in OWL-DL.

Backup-Slide: General structure of Actor Profile Ontology

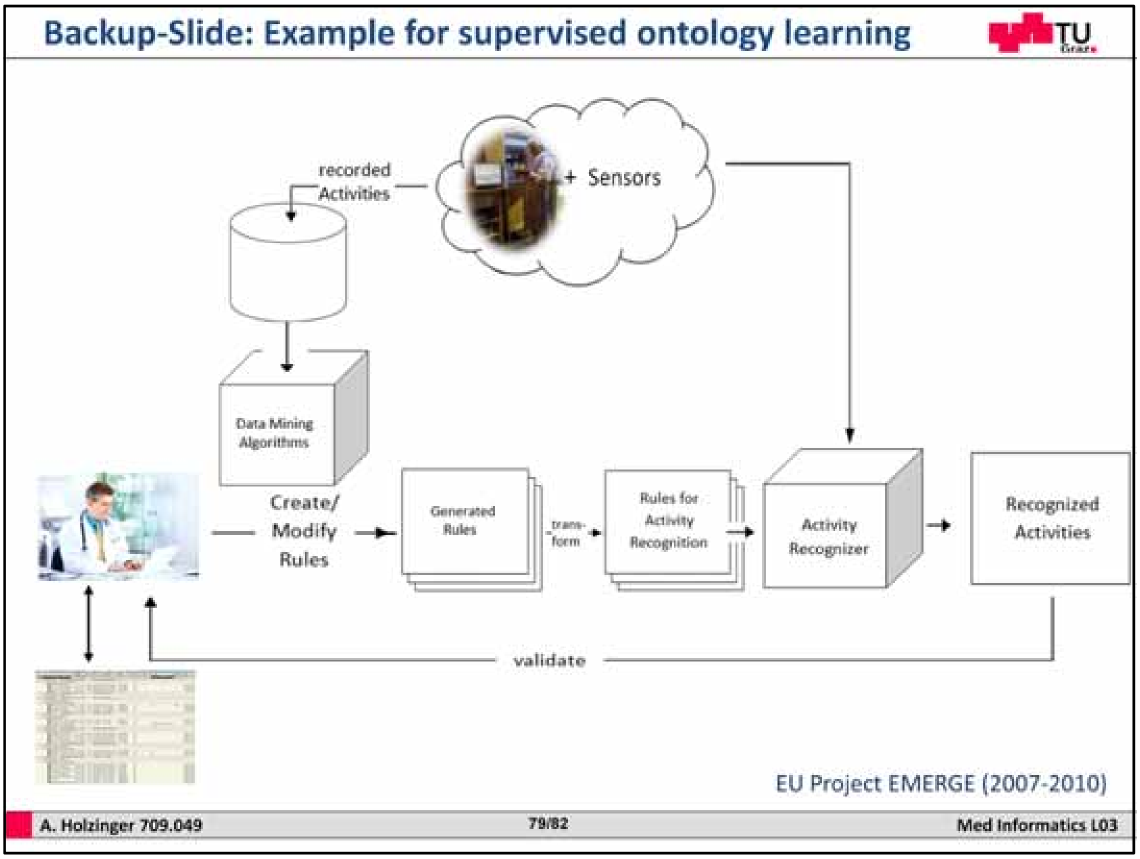


Valls et al. (2010)

A. Holzinger 709.049 77/82 Med Informatics L03



To solve this issues we made use of Semantic Web technologies. In order to provide a formal description of our concepts, terms and relationships within our knowledge domain we applied the Resource Description Framework (RDF) and the Web Ontology Language (OWL), particularly the OWL-DL (Description Logic). In the model you see the various layers from the low-level sensor layer up to the user-interaction layer and here you see the Environmental model, which includes the description of the physical objects (the sensor and location model); the Human-Capability model, which models medical expert knowledge, health parameters and interrelations, and the reaction and alarming schemes; the User model, containing the digital health record, end user specific settings, the social networks; these three models are integrated by the ontology based Information model, which describes the representation and semantics of the collected information objects, the semantic interoperability of components and is also responsible for the information quality.



At first the system must be trained
The first milestone was getting the medical knowledge into the system. At first we need a manual learning phase, where the underlying rules are defined manually by the medical professional. For this purpose we developed a user interface which allows an easy creation and editing of the rules. However, this kind of configuration entails a few disadvantages: The complexity grows up very fast with the number of possible underlying events (just as an example, for a proper toilet usage we have found 92 rules). Consequently, the adaptation of the rules to the individual behaviour and contexts is nearly impossible manually. Regarding these disadvantages, a promising approach was supervised learning: the system now gathers information about the typical user's behaviour during an initial learning phase automatically with feedback loop. All cases of concrete instances of the activities are now stored in a database. By application of data mining algorithms, characteristic sequences are automatically extracted and the rules for the activity recognizer automatically created. The medical professional is able to change all settings, to assure the best possible quality and to include the previous knowledge about the user e.g. from the patient record.

Backup-Slide: Expanding Queries with the MeSH Ontology

MeSH contains two organization files:

1) an alphabetic list with bags of synonymous and related terms, called records, and
2) a hierarchical organization of descriptors associated to the terms.

We consider that a term is a set of words (no word sequence order), that is:

$$t = \{w_1, \cdots, w_{|t|}\} \text{ where } w \text{ is a word}$$

A bag of terms is defined as:

$$b = \{t_1, \cdots, t_{|b|}\}$$

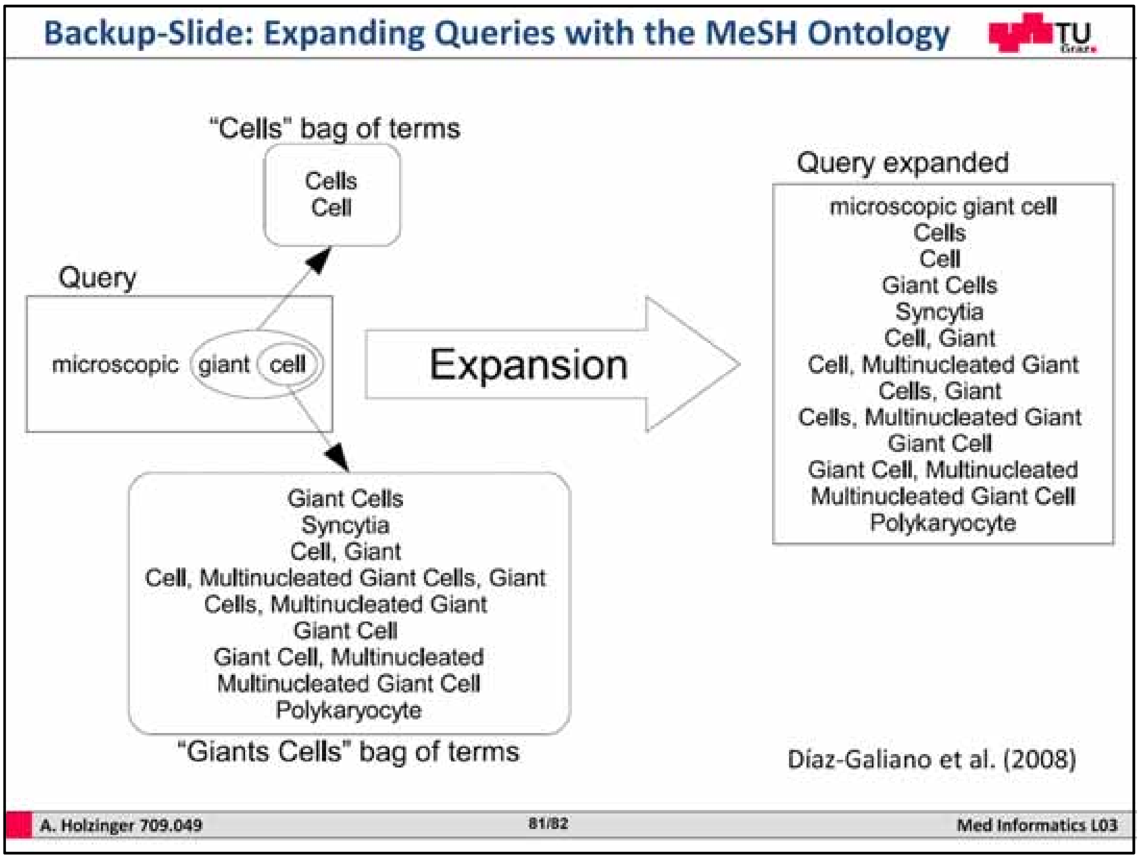
a term t exists in the query q $(t \in q)$ if:

$$\forall w_i \in t, \exists w_j \in q / w_i = w_j$$

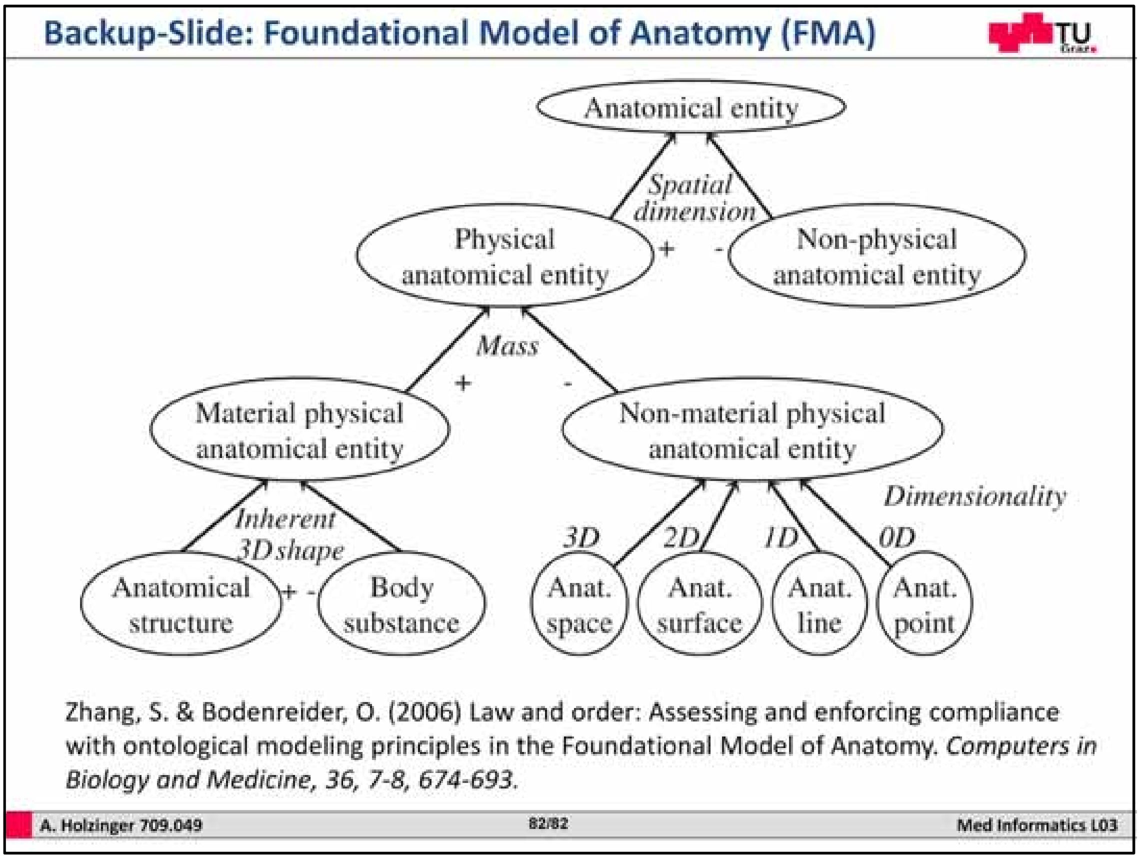
Therefore, if all the words of a term are in the query, we generate a new expanded query by adding all its bag of terms:

$$q \text{ is expanded with } b \text{ if } \exists t \in b / t \in q$$

Díaz-Galiano, M. et al. (2008) Integrating MeSH Ontology to Improve Medical Information Retrieval. In: Peters, C. et al. (Eds.) *Advances in Multilingual & Multimodal Information Retrieval, Lecture Notes in Computer Science 5152. Berlin, Heidelberg, New York, Springer, 601-606.*



In order to compare the words of a particular term to those of the query, all the words are put in lowercase and no stopword removal is applied. So as reduce the number of terms that could expand the query, we have only used those that are in A, C or E categories of MeSH (A: Anatomy, C: Diseases, E: Analytical, Diagnostic and Therapeutic Techniques and Equipment) [5].
Figure 1 shows an example of query expansion, with two terms found in the query and their bags of terms.



Top-level classes of the anatomy taxonomy of the FMA