What is an ontology? 
(And why should you care?)

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Outline

◆ What is an ontology?
  ● Definitions
  ● Examples of biomedical ontologies
  ● Applications of biomedical ontologies

◆ Why should you care?
  ● EHR-based research
  ● Promising ontology-enabled research
  ● Do not reinvent the wheel!
What is an ontology?
**Definition**

**Biomedical ontology**
- Artifact created for representing biomedical entities, their terms and their relations

**Ontology “spectrum”**
- Terminologies – focus on naming
- Controlled vocabularies / Thesauri – focus on use for indexing and retrieval (knowledge organization)
- Ontologies – support reasoning (logical definitions)

**Ontologies vs. knowledge bases**
- Definitional knowledge – mostly in ontologies
- Assertional knowledge – mostly in knowledge bases
Example Vocabulary

- Vocabulary – Collection of terms
  - epileptic fit grand mal
  - generalized seizures tonic clonic
  - generalized tonic clonic seizure
  - generalized tonic-clonic seizure
  - Grand mal
  - Grand mal epileptic fit
  - Grand mal fit
  - Grand mal seizure
  - haut mal
Example Relations

 Relations among entities

SNOMED Clinical Terms

- Clinical finding
- Neurological finding
- Seizure related finding
- Seizure
- Generalized seizure
- Tonic-clonic seizure
- Secondarily generalized seizures
- Head finding
- Finding of brain
- Finding of head and neck region
- Finding by site
- Central nervous system finding
- O/E - grand mal fit
- Grand mal seizure
Example: Logical definition

- Anticonvulsivant therapy (in SNOMED CT)

**Definition: Fully Defined as ...**

- is a
  - drug therapy
- has intent
  - therapeutic
- Group
  - method
    - administration - action
  - direct substance
    - anticonvulsivant
Example Knowledge base

- **Valproic acid (in NDF-RT)**
  - National Drug File – Reference terminology

<table>
<thead>
<tr>
<th>Kind</th>
<th>Subject</th>
<th>Predicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>VALPROIC ACID</td>
<td>isa</td>
</tr>
<tr>
<td>Disease</td>
<td>VALPROIC ACID</td>
<td>CI_with</td>
</tr>
<tr>
<td></td>
<td>VALPROIC ACID</td>
<td>CI_with</td>
</tr>
<tr>
<td></td>
<td>VALPROIC ACID</td>
<td>CI_with</td>
</tr>
<tr>
<td></td>
<td>VALPROIC ACID</td>
<td>may_treat</td>
</tr>
<tr>
<td></td>
<td>VALPROIC ACID</td>
<td>may_treat</td>
</tr>
<tr>
<td></td>
<td>VALPROIC ACID</td>
<td>may_treat</td>
</tr>
<tr>
<td></td>
<td>VALPROIC ACID</td>
<td>may_treat</td>
</tr>
<tr>
<td>Ingredient</td>
<td>VALPROIC ACID</td>
<td>has_ingredient</td>
</tr>
<tr>
<td>Mechanism of Action</td>
<td>VALPROIC ACID</td>
<td>has_MoA</td>
</tr>
<tr>
<td></td>
<td>VALPROIC ACID</td>
<td>has_MoA</td>
</tr>
<tr>
<td>Physiologic Effect</td>
<td>VALPROIC ACID</td>
<td>has_PE</td>
</tr>
</tbody>
</table>
3 examples of biomedical ontologies

- International Classification of Diseases (ICD)
- SNOMED Clinical Terms (SNOMED CT)
- Medical Subject Headings (MeSH)
International Classification of Diseases
ICD Characteristics (1)

- Current version: ICD-10
- Type: Classification
- Domain: Disorders
- Developer: World Health Organization (WHO)
- Funding: WHO

Availability
- Publicly available: No
- Repositories: UMLS and BioPortal

URL: [http://www.who.int/classifications/icd/en/](http://www.who.int/classifications/icd/en/)
ICD Characteristics (2)

◆ Number of
  ● Concepts: 12,318
  ● Terms: 1 per concept (tabular)

◆ Major organizing principles:
  ● Tree (single inheritance hierarchy)
  ● No explicit classification criteria
    ▪ Idiosyncratic inclusion/exclusion mechanism
  ● .8 slots for Not elsewhere classified (NEC)
  ● .9 slots for Not otherwise specified (NOS)

◆ Formalism: Proprietary format
ICD  Top level

ICD-10 Version:2010
- I Certain infectious and parasitic diseases
- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
- V Mental and behavioural disorders
- VI Diseases of the nervous system
- VII Diseases of the eye and adnexa
- VIII Diseases of the ear and mastoid process
- IX Diseases of the circulatory system
- X Diseases of the respiratory system
- XI Diseases of the digestive system
- XII Diseases of the skin and subcutaneous tissue
- XIII Diseases of the musculoskeletal system and connective tissue
- XIV Diseases of the genitourinary system
- XV Pregnancy, childbirth and the puerperium
- XVI Certain conditions originating in the perinatal period
- XVII Congenital malformations, deformations and chromosomal abnormalities
- XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- XIX Injury, poisoning and certain other consequences of external causes
- XX External causes of morbidity and mortality
- XXI Factors influencing health status and contact with health services
- XXII Codes for special purposes
Epilepsy in ICD-10 (1)

Chapter VI
Diseases of the nervous system
(G00-G99)

Episodic and paroxysmal disorders
(G40-G47)

<table>
<thead>
<tr>
<th>G40</th>
<th>Epilepsy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Excl.:</td>
</tr>
<tr>
<td></td>
<td>Landau-Kleffner syndrome (E80.3)</td>
</tr>
<tr>
<td></td>
<td>seizure (convulsive) NOS (R56.8)</td>
</tr>
<tr>
<td></td>
<td>status epilepticus (G41.8)</td>
</tr>
<tr>
<td></td>
<td>Todd's paralysis (G83.8)</td>
</tr>
</tbody>
</table>

G40.0 Localization-related (focal)(partial) idiopathic epilepsy and epileptic syndromes with seizures of localized onset
- Benign childhood epilepsy with centrotemporal EEG spikes
- Childhood epilepsy with occipital EEG paroxysms

G40.1 Localization-related (focal)(partial) symptomatic epilepsy and epileptic syndromes with simple partial seizures
- Attacks without alteration of consciousness
- Simple partial seizures developing into secondarily generalized seizures

G40.2 Localization-related (focal)(partial) symptomatic epilepsy and epileptic syndromes with complex partial seizures
- Attacks with alteration of consciousness, often with automatisms
- Complex partial seizures developing into secondarily generalized seizures

G40.3 Generalized idiopathic epilepsy and epileptic syndromes
Epilepsy in ICD-10 (2)

G40.4 Other generalized epilepsy and epileptic syndromes
Epilepsy with:
- myoclonic absences
- myoclonic-astatic seizures
Infantile spasms
- Lennox-Gastaut syndrome
Salaam attacks
Symptomatic early myoclonic encephalopathy
West's syndrome

G40.5 Special epileptic syndromes
Epilepsia partialis continua [Kozhevnikof]
Epileptic seizures related to:
- alcohol
- drugs
- hormonal changes
- sleep deprivation
- stress

Use additional external cause code (Chapter XX), if desired, to identify drug, if drug-induced.

G40.6 Grand mal seizures, unspecified (with or without petit mal)
G40.7 Petit mal, unspecified, without grand mal seizures
G40.8 Other epilepsy
Epilepsies and epileptic syndromes undetermined as to whether they are focal or generalized

G40.9 Epilepsy, unspecified
Epileptic:
- convulsions NOS
- fits NOS
- seizures NOS
SNOMED Clinical Terms
SNOMED CT Characteristics (1)

- Current version: July 31, 2012 (2 annual releases)
- Type: Reference terminology / ontology
- Domain: Clinical medicine
- Developer: IHTSDO
- Funding: IHTSDO
- Availability
  - Publicly available: Yes* (in member countries)
  - Repositories: UMLS
- URL: http://www.ihtsdo.org/
SNOMED CT Characteristics (2)

- **Number of**
  - Concepts: ~300,000 active concepts (July 31, 2012)
  - Terms: ~1M active “descriptions”

- **Major organizing principles:**
  - Utility for clinical medicine (e.g., assertional + definitional knowledge)
  - Model of meaning (incomplete)
  - Rich set of associative relationships
  - Small proportion of defined concepts (many primitives)

- **Formalism:** Description logics (KRSS)
SNOMED CT Top level
Epilepsy in SNOMED CT
Medical Subject Headings (MeSH)
MeSH Characteristics (1)

- Current version: 2012 (yearly releases)
- Type: Thesaurus / Controlled vocabulary
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (Library Operations)

Availability
- Publicly available: Yes
- Repositories: UMLS / NCBO BioPortal

URL: http://www.nlm.nih.gov/mesh/
MeSH Characteristics (2)

◆ Number of
  ● Terms: 7.5 per descriptor

◆ Major organizing principles:
  ● Descriptor + entry terms
  (also: Qualifiers, Supplementary concepts)
  ● Thesaurus relations (RB/RN/RO)

◆ Formalism: Thesaurus / Proprietary XML DTD
MeSH Top level

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 and 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]
# MeSH Example (terms)

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Epilepsy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Number</td>
<td>C10.228.140.490</td>
</tr>
<tr>
<td>Annotation</td>
<td>GEN or unspecified; prefer specifics</td>
</tr>
<tr>
<td>Scope Note</td>
<td>A disorder characterized by recurrent episodes of paroxysmal brain dysfunction due to a sudden, disorderly, and excessive neuronal discharge. Epilepsy classification systems are generally based upon: (1) clinical features of the seizure episodes (e.g., motor seizure), (2) etiology (e.g., post-traumatic), (3) anatomic site of seizure origin (e.g., frontal lobe seizure), (4) tendency to spread to other structures in the brain, and (5) temporal patterns (e.g., nocturnal epilepsy). (From Adams et al., Principles of Neurology, 6th ed, p313)</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Aura</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Awakening Epilepsy</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Epilepsy, Cryptogenic</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Epileptic Seizures</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Seizure Disorder</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Seizures, Epileptic</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Single Seizure</td>
</tr>
<tr>
<td>See Also</td>
<td>Seizures</td>
</tr>
<tr>
<td>Allowable Qualifiers</td>
<td>BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IN</td>
</tr>
<tr>
<td>Date of Entry</td>
<td>19990101</td>
</tr>
<tr>
<td>Unique ID</td>
<td>D004827</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of Substance</th>
<th>Epileptic encephalopathy, Lennox-Gastaut type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record Type</td>
<td>C</td>
</tr>
<tr>
<td>Registry Number</td>
<td>0</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Encephalopathy of childhood</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Lennox-Gastaut syndrome</td>
</tr>
<tr>
<td>Heading Mapped to</td>
<td>*Intellectual Disability</td>
</tr>
<tr>
<td>Heading Mapped to</td>
<td>*Spasms, Infantile</td>
</tr>
<tr>
<td>Frequency</td>
<td>65</td>
</tr>
<tr>
<td>Date of Entry</td>
<td>20100825</td>
</tr>
<tr>
<td>Unique ID</td>
<td>C535500</td>
</tr>
</tbody>
</table>
MeSH Example (hierarchies)

Nervous System Diseases [C10]
  Central Nervous System Diseases [C10.228]
    Brain Diseases [C10.228.140]
    Epilepsy [C10.228.140.490]

  Epilepsies, Myoclonic [C10.228.140.490.250] +
  Epilepsies, Partial [C10.228.140.490.360] +
  Epilepsy, Benign Neonatal [C10.228.140.490.370]
  ▶ Epilepsy, Generalized [C10.228.140.490.375]
    Epilepsy, Absence [C10.228.140.490.375.260]
    Epilepsy, Tonic-Clonic [C10.228.140.490.375.290]
    Spasms, Infantile [C10.228.140.490.375.760]
  Epilepsy, Post-Traumatic [C10.228.140.490.380]
  Epilepsy, Reflex [C10.228.140.490.450]
  Landau-Kleffner Syndrome [C10.228.140.490.535]
  Seizures [C10.228.140.490.631]
  Seizures, Febrile [C10.228.140.490.650]
  Status Epilepticus [C10.228.140.490.690] +
Applications of biomedical ontologies

3 major categories of use

- **Knowledge management** (indexing and retrieval of data and information, access to information, mapping among ontologies)
- **Data integration**, exchange and semantic interoperability
- **Decision support and reasoning** (data selection and aggregation, decision support, natural language processing applications, knowledge discovery).

[Bodenreider, YBMI 2008]
Applications Knowledge management

◆ Source of vocabulary and definitions for
  - Annotation (e.g., text mining)
  - Indexing (and information retrieval)
  - Coding (clinical records, registries)

◆ Source of mapping across ontologies
  - Correspondence between terms across ontologies
  - Terminology integration systems
    - Unified Medical Language System
    - BioPortal
Applications  Interoperability

◆ Exchange of standardized data
  ● Use of standard vocabularies and protocols
  ● E.g., HL7 messages, clinical documents (CDA)

◆ Data integration
  ● Clinical data warehouses (for translational research)
    ■ Data standardized while loading
  ● Query translation against federated data repositories
    ■ Supported by ontologies
Applications  Decision support / reasoning

◆ Hierarchical information in ontologies helps bridge across levels of granularity
  - CDS rule expressed at the level of a pharmacologic class
    - E.g., interactions between anti-convulsivants and other drugs
  - Cohort selection
    - Retrieve all patients admitted for seizure
  - Hypothesis generation / Knowledge discovery
    - Aggregate relations to increase statistical power
    (e.g., aggregate drugs at the level of pharmacologic class in relation to adverse events)
Why should you care?
3 reasons why you should care

- EHR-based research
- Promising ontology-enabled research
- Do not reinvent the wheel!
EHR-based research

◆ Clinical data warehouses
  ● Secondary use of clinical data
  ● Provide large amounts of data
    ▪ Much larger than regular clinical trials
  ● Integration of clinical data across institutions
    ▪ When standardized

◆ Enable new forms of clinical research
  ● Translational research
    ▪ Integration with genomic data
  ● Study of adverse events based on observational data
    ▪ Post-marketing surveillance of drugs
  ● “Learning Healthcare System”
    ▪ EHR data become knowledge
Promising ontology-enabled research

Basic ingredients

- Vocabulary from ontologies is used to standardize mentions of biomedical entities in a corpus (MEDLINE articles, EHR data)
  - Term extraction systems (natural language processing)
- Ontological relations are used to aggregate semantically similar datapoints in order to increase statistical power
  - Clustering based on semantic similarity
- Data mining techniques are used to identify novel, salient relations among biomedical entities
Do not reinvent the wheel!

- Leverage existing ontologies (for annotation, aggregation, etc.)
  - Jump on the “standards” bandwagon
- Ontology development is difficult and expensive
  - Be clinicians / researchers; collaborate with ontologists
- There are already many clinical ontologies available
  - Basic research must follow the same standards as clinical research to become “translational”
References

- Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67-79. PubMed PMID: 18660879; PubMed Central PMCID: PMC2592252

Resources

◆ Unified Medical Language System (UMLS)
  ● National Library of Medicine
  ● https://uts.nlm.nih.gov/

◆ BioPortal
  ● National Center for Biomedical Ontology
  ● http://bioportal.bioontology.org/

◆ Open Biomedical Ontology
  ● OBO Community
  ● http://www.obofoundry.org/
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA