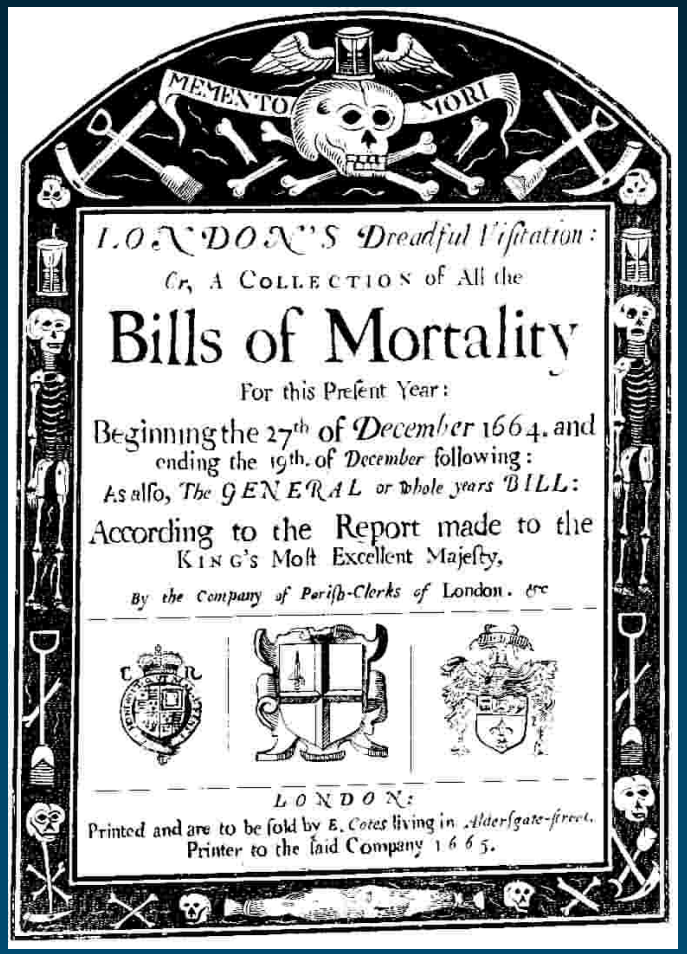
Public Health Informatics Fellowship Program
March 12, 2010

Ontologies and Biosurveillance

Olivier Bodenreider

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

London Bills of Mortality



MEMENTO MORI

LONDON'S Dreadful Visitation:

Or, A COLLECTION of All the

Bills of Mortality

For this Present Year:

Beginning the 27th of December 1664. and ending the 19th. of December following:

As also, The GENERAL or whole years BILL:

According to the Report made to the KING's Most Excellent Majesty,

By the Company of Parish-Clerks of London. &c

LONDON:

Printed and are to be sold by E. Cotes living in Aldersgate-street. Printer to the said Company 1665.

A generall Bill for this present year, ending the 19 of December 1665. according to the Report made to the KINGS most Excellent Majesty.

By the Company of Parish Clerks of London, &c.

The Diseases and Casualties this year.

Abortive and Stilborne	617	Executed	21	Palsie	30
Aged	1545	Flux and Small Pox	655	Plague	68596
Ague and Peaver	5257	Found dead in streets, fields, &c.	2[illegible]	Plannet	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	13
Bedrid	10	Frighted	23	Poysoned	1
Blasted	5	Gout and Sciatica	27	Quinsie	35
Bleeding	16	Grief	46	Rickets	157
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1238	Riling of the Lights	397
Burnt and Scalded	8	Hangd & made away themselves	7	Rupture	34
Calenture	3	Headmouldshot & Mouldfallen	14	Scurvy	1[illegible]7
Cancer, Gangrene and Fistula	56	Jaundies	1[illegible]	Shingles and Swine pox	2
Canker, and Thrush	1[illegible]	Impostume	227	Stores, Ulcers, broken and bruised	
Childbed	625	Kild by severall accidents	46	Limbs	82
Christomes and Infants	1258	Kings Evill	86	Spleen	14
Cold and Cough	68	Leprosie	2	Spotted Fever and Purples	1929
Collick and Winde	134	Lethargy	14	Stopping of the stomack	33[illegible]
Consumption and Tissick	48[illegible]8	Livergrown	2[illegible]	Stone and Strangury	[illegible]8
Convulsion and Mother	2056	Meagrom and Headach	1[illegible]	Surfet	[illegible]
Distracted	[illegible]	Measles	7	Teeth and Worms	2614
Drouse and Tympany	1476	Murthered and Shot	9	Vomiting	[illegible]
Drowned	[illegible]	Overlaid & Starved	45	Venn	[illegible]

Christned: Males 5114, Females 4853, In all 9967

Buried: Males 48568, Females 48737, In all 97[illegible]

Of the Plague 68596

Increased in the Burials in the 130 Parishes and at the Pest-house this year 79[illegible]

Increased of the Plague in the 130 Parishes and at the Pest-house this year 68590

Limitations of existing classifications

"The advantages of a uniform statistical nomenclature, however imperfect, are so obvious, that it is surprising no attention has been paid to its enforcement in Bills of Mortality. Each disease has, in many instances, been denoted by three or four terms, and each term has been applied to as many different diseases: vague, inconvenient names have been employed, or complications have been registered instead of primary diseases. The nomenclature is of as much importance in this department of inquiry as weights and measures in the physical sciences, and should be settled without delay."

– William Farr

First annual report.

London, Registrar General of England and Wales, 1839, p. 99.

Outline

- Biomedical ontologies
  - What they are
  - What they are for
- Ontologies and biosurveillance
  - Support for text mining
  - Controlled vocabulary
  - Data aggregation
  - Data integration
  - Reasoning
- Ontology in action in a biosurveillance system – BioCaster

Biomedical ontologies

What they are

Overview

[J. Cimino, YBMI 2006]

- Structural perspective
  - What are they (vs. what are they for)?
- "High-impact" biomedical ontologies
  - International Classification of Diseases (ICD)
  - Logical Observation Identifiers, Names and Codes (LOINC)
  - SNOMED Clinical Terms
  - Foundational Model of Anatomy
  - Gene Ontology
  - RxNorm
  - Medical Subject Headings (MeSH)
  - NCI Thesaurus
  - Unified Medical Language System (UMLS)

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 [ICD9-CM in NCBO BioPortal]
- URL: 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

Chapter	Blocks	Title
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D48	Neoplasms
III	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	E00-E90	Endocrine, nutritional and metabolic diseases
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous tissue
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Pregnancy, childbirth and the puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

ICD Example

- Idiosyncratic inclusion/exclusion criteria

E10 **Insulin-dependent diabetes mellitus**

[See before E10 for subdivisions]

Includes: diabetes (mellitus):
- brittle
- juvenile-onset
- ketosis-prone
- type I

Excludes: diabetes mellitus (in):
- malnutrition-related (E12.-)
- neonatal (P70.2)
- pregnancy, childbirth and the puerperium (O24.-)

glycosuria:
- NOS (R81)
- renal (E74.8)

impaired glucose tolerance (R73.0)

postsurgical hypoinsulinaemia (E89.1)

ICD Example

- Not elsewhere classified (NEC)
- Not otherwise specified (NOS)

E84	**Cystic fibrosis**
	Includes: mucoviscidosis
E84.0	**Cystic fibrosis with pulmonary manifestations**
E84.1	**Cystic fibrosis with intestinal manifestations**
	Meconium ileus+ (P75*)
	Excludes: meconium obstruction in cases where cystic fibrosis is known not to be present (P76.0)
E84.8	**Cystic fibrosis with other manifestations**
	Cystic fibrosis with combined manifestations
E84.9	**Cystic fibrosis, unspecified**

Logical Observation Identifiers, Names and Codes (LOINC)

LOINC Characteristics (1)

- Current version: 2.27 (July 2009)
- Type: Controlled terminology*
- Domain: Laboratory and clinical observations
- Developer: Regenstrief Institute
- Funding: NLM
- Availability
  - Publicly available: Yes
  - Repositories: UMLS
- URL: www.regenstrief.org/loinc/loinc.htm

LOINC Characteristics (2)

- Number of
  - Concepts: 50k active codes (2.18)
  - Terms: n/a*
- Major organizing principles:
  - No hierarchical structure among the main codes
  - 6 axes
    - Component (analyte [+ challenge] [+ adjustments])
    - Property
    - Timing
    - System
    - Scale
    - [Method]
- Formalism: “DL-like”

LOINC Example

- *Sodium:SCnc:-Pt:Ser/Plas:Qn*
  [the molar concentration of sodium is measured in the plasma (or serum), with quantitative result]

Axis	Value
Component	Sodium
Property	SCnc – Substance Concen-tration (per volume)
Timing	Pt – Point in time (Random)
System	Ser/Plas – Serum or Plasma
Scale	Qn – Quantitative
Method	--

SNOMED Clinical Terms

SNOMED CT Characteristics (1)

- Current version: January 31, 2009 (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: 311,313 active concepts (Jan. 31, 2008)
  - Terms: 794,061 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

Hierarchy | Subtype hierarchy

138875005	SNOMED CT Concept
362981000	qualifier value
106237007	linkage concept
370115009	special concept
48176007	social context
419891008	record artifact
363787002	observable entity
308916002	environment or geographical location
123038009	specimen
254291000	staging and scales
123037004	body structure
272379006	event
78621006	physical force
404684003	clinical finding
260787004	physical object
410607006	organism
71388002	procedure
373873005	pharmaceutical / biologic product
243796009	situation with explicit context
105590001	substance

SNOMED CT Example

Hierarchy | Subtype hierarchy

- 27010001 partial excision of large intestine
- 8613002 operation on appendix
  - 80146002 appendectomy
    - 82730006 incidental appendectomy
    - 49438003 appendectomy with drainage
    - 174036004 emergency appendectomy
    - 174045003 interval appendectomy
    - 6025007 laparoscopic appendectomy
    - 235313004 non-emergency appendectomy
    - 235314005 inversion appendectomy
    - 1299000 excision of appendiceal stump

Definition: Fully defined by ...

- is a
  - D partial excision of large intestine
  - D operation on appendix
- *Group*
  - method
    - D excision - action
  - procedure site - Direct
    - D appendix structure
- *Qualifiers*
  - access
    - surgical access values
  - priority
    - priorities

appendectomy - Definition

Concept Status: **Current**

- *Descriptions*
  - F appendectomy (procedure)
  - P appendectomy
  - S excision of appendix
  - U appendicectomy

Codes

- *Original SnomedId : P1-57450*
- *Read Code (Ctv3Id) : X20Wz*

RxNorm

RxNorm Characteristics (1)

- Current version: July 6, 2009 (monthly releases)
- Type: Controlled terminology
- Domain: Drug names
- Developer: NLM
- Funding: NLM
- Availability
  - Publicly available: Yes*
  - Repositories: UMLS
- URL: http://www.nlm.nih.gov/research/umls/rxnorm/

RxNorm Characteristics (2)

- Number of
  - Concepts: 93k (June 2008)
  - Terms: 105k
- Major organizing principles:
  - Generic vs. brand
  - Combinations of Ingredient / Form / Dose
  - No hierarchical structure
  - Links to all major US drug information sources
  - No clinical information
- Formalism: UMLS RRF format

RxNorm Normalized form

Strength **Ingredient** **Dose form**

4mg/ml Fluoxetine Oral Solution

Strength **Ingredient**

Semantic clinical drug component

Ingredient **Dose form**

Semantic clinical drug form

Strength **Ingredient** **Dose form**

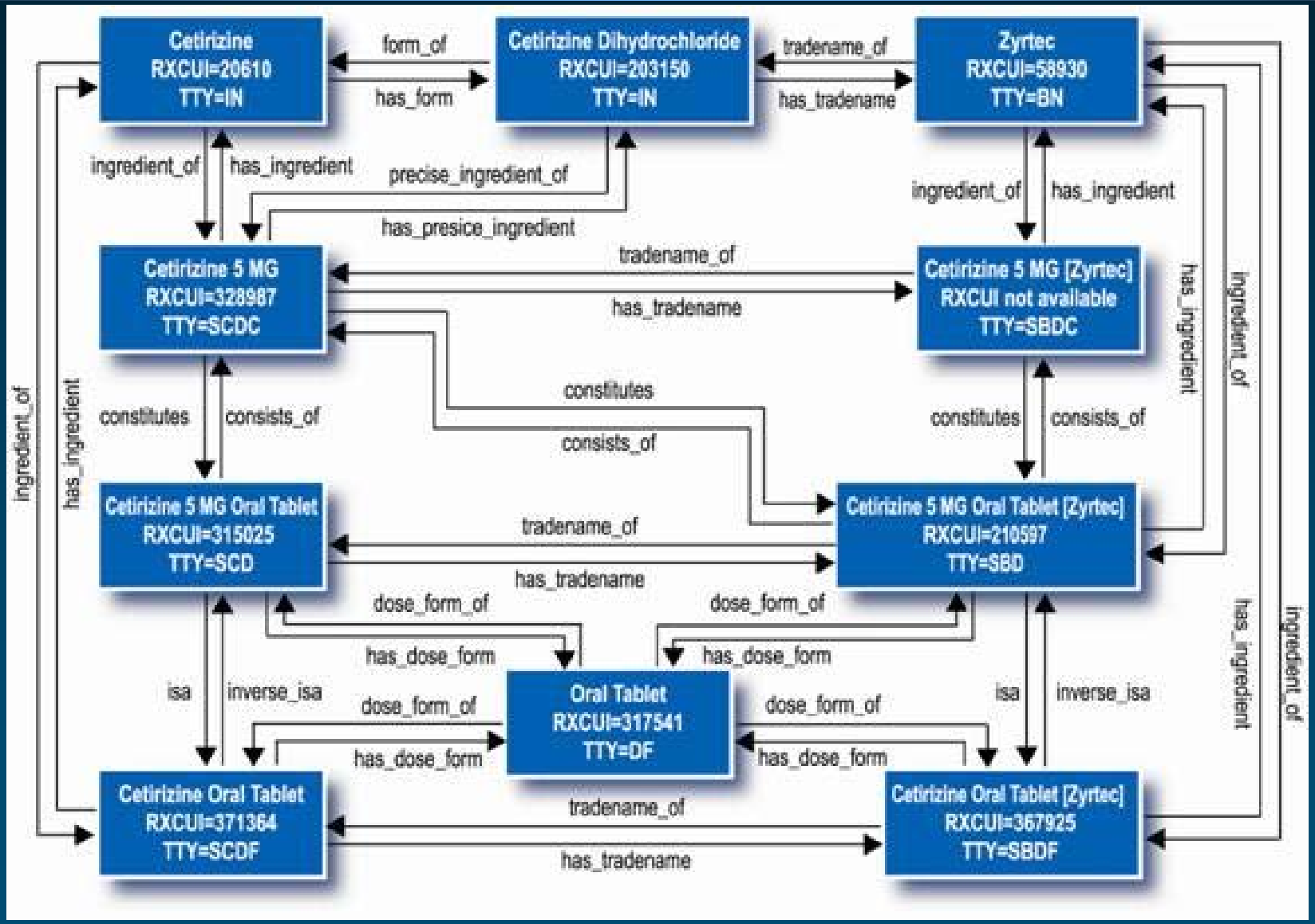
Semantic clinical drug

Rx Norm Generic *vs.* Brand

- Generic
  - Ingredient (IN)
  - Clinical drug form (SCDF)
  - Clinical drug component (SCDC)
  - Clinical drug (SCD)
- Brand
  - Brand name (BN)
  - Branded drug form (SBDF)
  - Branded drug component (SBDC)
  - Branded drug (SBD)

tradename_of

RxNorm Relations among drug entities



Recap

Name	Scope	# concepts	Median	Subs. Hier	Version
SNOMED CT	Clinical medicine (patient records)	310,314	2	yes	July 31, 2007
LOINC	Clinical observations and laboratory tests	46,406	3	no	Version 2.21 (no “natural language” names)
FMA	Human anatomical structures	~72,000	?	yes	(not yet in the UMLS)
Gene Ontology	Functional annotation of gene products	22,546	1	yes	Jan. 2, 2007
RxNorm	Standard names for prescription drugs	93,426	1	no	Aug. 31, 2007
NCI Thesaurus	Cancer research, clinical care, public information	58,868	2	yes	2007_05E
ICD-10	Diseases and conditions (health statistics)	12,318	1	no	1998 (tabular)
MeSH	Biomedicine (descriptors for indexing the literature)	24,767	5	no	Aug. 27, 2007
UMLS .	Terminology integration in the life sciences	1,4 M	2	n/a	2007AC (English only)

[Bodenreider, YBMI 2008]

Biomedical ontologies

What they are for

Overview

[Bodenreider, YBMI 2008]

- Functional perspective
  - What are they for (vs. what are they)?
- "High-impact" 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).

Biomedical ontologies

Needs for biosurveillance

Needs for biosurveillance

- Support for text mining
- Controlled vocabulary
- Aggregation
- Data integration
- Reasoning

Support for text mining

- Lexical resources
  - Identify mentions in text
    - Lexical variants
- Terminological resources
  - Identify concepts
    - Synonyms
- Ontological resources
  - Identify relations, Semantic interpretation
    - Domain knowledge

Controlled vocabulary

- Coded information
  - Storage
  - Processing
- Standardize
  - Definitions
  - Usage

A generall Bill for this present year, ending the 19 of *December* 1665. according to the Report made to the KINGS most Excellent Majesty.

By the Company of Parish Clerks of *London*, &c.

The Diseases and Casualties this year.

Abortive and Stilborne	617	Executed	21	Palsie	30
Aged	1545	Flux and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in streets, fields, &c.	20	Plannet	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	15
Bedrid	10	Frighted	23	Poysoned	1
Blasted	5	Gout and Sciatica	27	Quinsie	35
Bleeding	16	Grief	46	Rickets	157
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Riling of the Lights	197
Burnt and Scalded	8	Hangd & made away themselves	7	Rupture	34
Calenture	3	Headmouldshot & Mouldfallen	14	Scurvy	107
Cancer, Gangrene and Fistula	56	Jaundies	120	Shingles and Swine pox	2
Canker, and Thrush	121	Impostume	227	Sores, Ulcers, broken and bruised	
Childbed	625	Kild by severall accidents	46	Limbs	82
Chrisomes and Infants	1258	Kings Evill	86	Spleen	14
Cold and Cough	68	Leprosie	2	Spotted Feaver and Purples	1929
Collick and Winde	134	Lethargy	14	Stopping of the stomack	332
Consumption and Tissick	4808	Livergrown	21	Stone and Strangury	98
Convulsion and Mother	2036	Meagrom and Headach	12	Surfet	[illegible]
Distracted	5	Measles	7	Teeth and Worms	2614
Drouse and Tympany	1476	Murthered and Shot	9	Vomiting	71
Drowned	52	Overlaid & Starved	45	Venn	7

Christned: Males	5114	Buried: Males	48568	Of the Plague	68596
Christned: Females	4853	Buried: Females	48737		
Christned: In all	9967	Buried: In all	97306		

Increased in the Burials in the 130 Parishes and at the Pest-house this year — 79009

Increased of the Plague in the 130 Parishes and at the Pest-house this year — 68590

Data aggregation

- Granularity mismatch
  - Data recorded
  - Data needed for making decisions
- Aggregation along hierarchies
  - Subsumption hierarchies (isa)
  - Ad hoc linearizations (e.g., ICD for mortality / morbidity)

Data integration

- Datasets annotated in reference to multiple ontologies
- Establish correspondence between equivalent concepts across ontologies (and datasets)
- Role of terminology integration systems
  - UMLS, RxNorm
  - NCBO ontology services

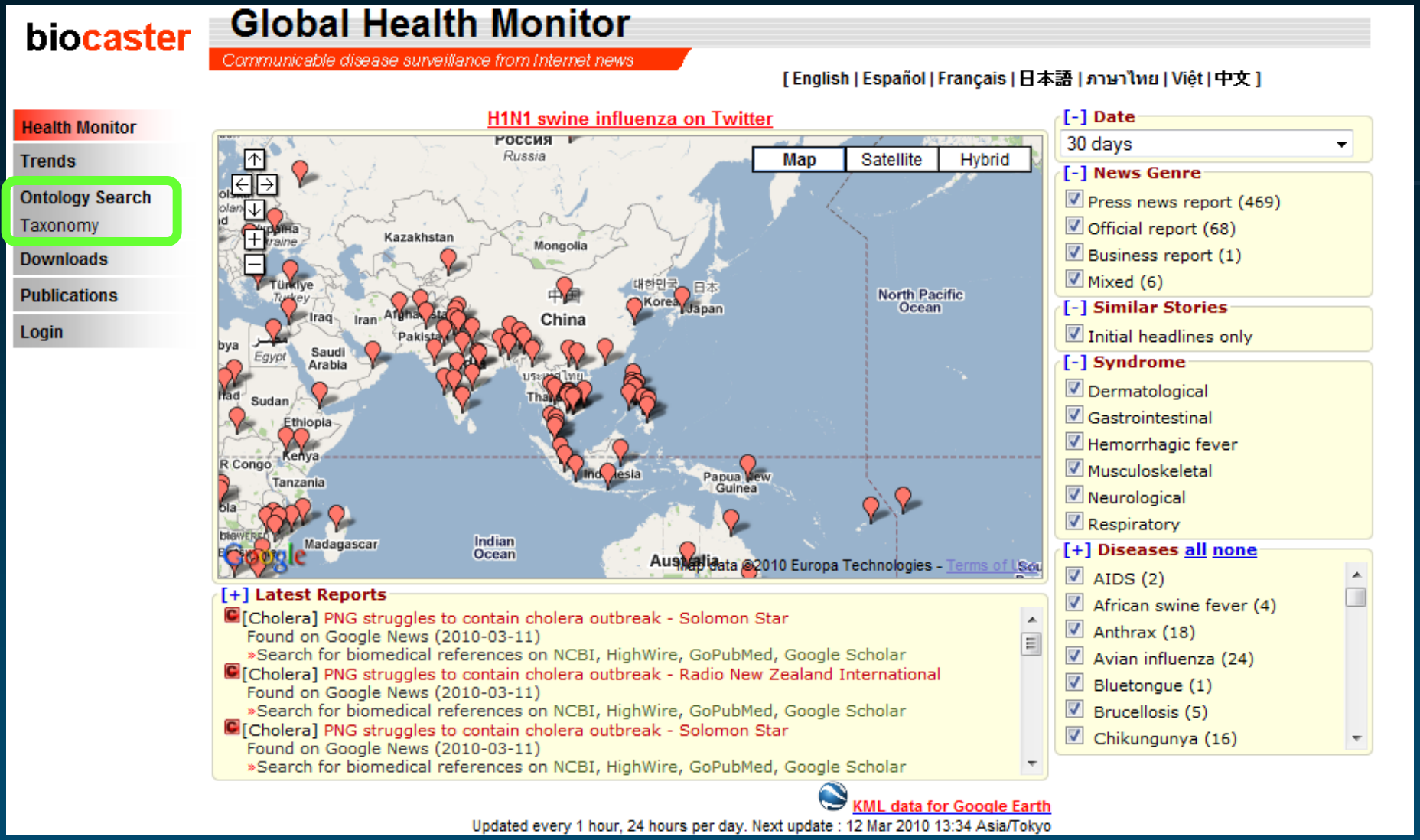
Reasoning

- Description Logics
- Reasoning services (DL classifiers)
  - Instance classification

Ontology in action in a biosurveillance system

BioCaster

http://biocaster.org/



http://biocaster.org/

BioCaster

- "Ontology-based text mining system for detecting and tracking the distribution of infectious disease outbreaks from linguistic signals on the Web"
- In operation since 2006
- Scans 1700 RSS feeds
- 4 stages
  - topic classification
  - named entity recognition (NER)
  - disease/location detection
  - event recognition

BioCaster

- BioCaster ontology
  - Vocabulary for Named Entity Recognition
    - Bridges between layman's terms and biomedical concepts
    - Multi-lingual (8 languages)
  - Knowledge about
    - Infectious diseases (e.g., causal agent, manifestations, associated syndrome)
  - Mappings to reference terminologies

http://biocaster.nii.ac.jp/index.php?mod=ontology&task=view&type=DISEASE&id=14121

BioCaster ontology

Taxonomy

- AbstractEntity
  - Attribute
    - BiologicalAttribute
      - CONDITION
      - SYMPTOM
        - Asymptomatic
        - EarNoseThroatOralCavitySymptom
        - FeverSymptom
        - **GastrointestinalTractSymptom**
          - Abdominal pain
          - Blood in stool
          - Blood in vomit
          - Coffee ground vomiting
          - Constipation
          - Diarrhea
          - Dysentery
          - Gastrointestinal bleeding
          - Gastrointestinal inflammation
          - Hemorrhagic diarrhea
          - Loss of appetite
          - Nausea
          - Vomit
        - GenitourinarySymptom
        - HematologicalSymptom

BioCaster ontology Cholera

Concept Details

DISEASE

Identifiers

Name	Cholera
Code	DISEASE_3
Definition	Cholera is an acute infectious gastrointestinal disease, caused by consuming the water or food that is contaminated with the bacterium Vibrio cholerae.
Preferred term	

Information about this concept

Synonyms	en: Red death (DISEASE) en: Cholera (DISEASE) en: Asiatic chorela (DISEASE) fr: Choléra (DISEASE) jp: コレラ (DISEASE) ko: 콜레라 (DISEASE) zh: 霍乱 (DISEASE) es: Cólera (DISEASE) th: อหิวาตกโรค (DISEASE) vi: Bệnh dịch tả (DISEASE) vi: Bệnh tả (DISEASE)

BioCaster ontology Cholera

Causal agent	Vibrio cholerae (BACTERIUM)
Symptoms	Vomit (GastrointestinalTractSymptom) Nausea (GastrointestinalTractSymptom) Diarrhea (GastrointestinalTractSymptom) Circulatory collapse (HematologicalSymptom) Acidosis (HematologicalSymptom) Dehydration (MetabolismSymptom) Apathy (PsychologicalSymptom)
Similar to	
Associated syndrome	Gastrointestinal syndrome (SYNDROME)

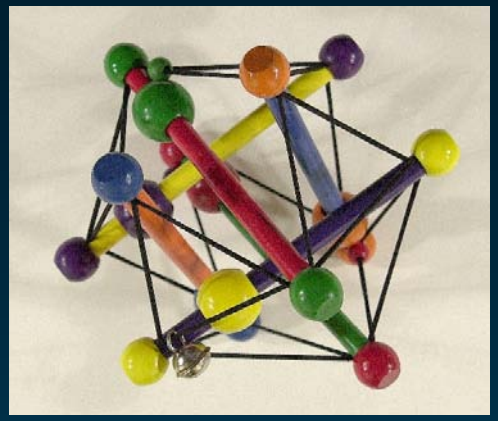
Super concepts

Subconcepts

BioCaster ontology Cholera

External links

- ICD-10: Cholera: A00
- ICD-9: Cholera: 001
- LOINC: Cholera due to Vibrio cholerae (disorder): DE-11601
- LOINC: Cholera (disorder): DE-11600
- MedDRA: Cholera, unspecified [10008634]: 10008634
- MedDRA: Cholera due to Vibrio cholerae el tor [10008633]: 10008633
- MedDRA: Cholera due to Vibrio cholerae [10008632]: 10008632
- MedDRA: Cholera [10008631]: 10008631
- MeSH: Cholera: C01.252.400.959.347
- SNOMED CT: Cholera (disorder): 63650001
- Wikipedia: Cholera



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

References Ontologies

- Bodenreider O, Stevens R.
  Bio-ontologies: current trends and future directions.
  Brief Bioinform. 2006 Sep;7(3):256-74.
- Cimino JJ, Zhu X.
  The practical impact of ontologies on biomedical informatics.
  Yearb Med Inform. 2006:124-35.
- Bodenreider O.
  Biomedical ontologies in action: role in knowledge management, data integration and decision support.
  Yearb Med Inform. 2008:67-79.

References BioCaster

- Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi
  BioCaster: detecting public health rumors with a Web-based text mining system
  Bioinformatics. 2008 December 15; 24(24): 2940–2941.
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2639299/