Health-e-Child Final Conference

Sestri Levante, Italy
April 24, 2010

Present and Future of Biomedical Ontologies

Olivier Bodenreider

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

Outline

- A few legitimate questions
  - What do you mean by *ontology*?
  - Why a talk on biomedical ontologies at the Health-e-Child conference?
- Biomedical ontologies
  - (A quick look at) The past
  - The present
  - The future

[Bodenreider, Brief Bioinf 2006]

What do you mean by *ontology*?

Ontology vs. other artifacts

- Ontology
  - Defining types of things and their relations
- Terminology
  - Naming things in a domain
- Thesaurus
  - Organizing things for a given purpose
- Classification
  - Placing things into (arbitrary) classes
- Knowledge bases
  - Assertional vs. definitional knowledge

Ontology vs. other artifacts (revisited)

- Lexical and terminological resources
  - Mostly collections of names for biomedical entities
  - Often have some kind or hierarchical organization (e.g., relations)
- Ontological resources
  - Mostly collections of relations among biomedical entities
  - Sometimes also collect names

"Ontological spectrum"

Why a talk on biomedical ontologies at the Health-e-Child conference?

Ontology in HeC presentations

- Nomenclatures
  - O. Milenasi
    - International Paediatric and Congenital Cardiac Codes
    - Normalization efforts (in Europe and the EU)
    - Harmonization of the two nomenclatures
- Ontologies
  - A. Tsymbal
    - Gene Ontology, KEGG
    - Reasoning based on ontologies (e.g., semantic similarity)
  - A. Everett
    - Abstract clinical information from patient records
    - Facilitate the recruitment of patients for clinical trials

HeC and ontology

- HeC
  - Outcomes, diagnoses, procedures
  - Personalized medicine
- Sharing information requires normalization
  - Among healthcare practitioners
  - Through clinical research databases
    - Evidence-based medicine
    - Comparative effectiveness
- Analyzing information requires aggregation
  - Compensate for differences in granularity

Genotype vs. phenotype

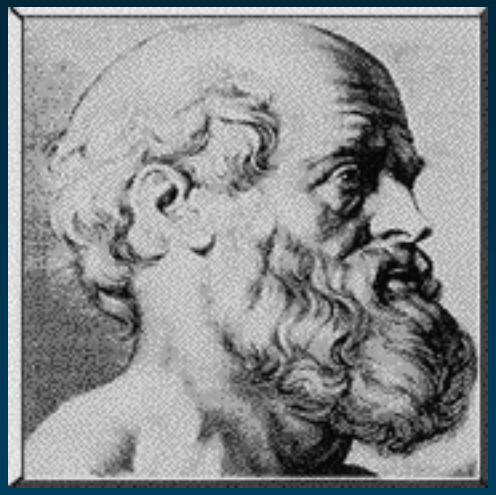
- Genotype information
  - "Exact" measurement (+ context)
  - Can be easily analyzed through mathematical models
    - Micro-arrays
    - Sequence similarity
    - SNP patterns
- Phenotype information
  - Results from human observation (+ interpretation / context)
  - Requires normalization
  - May require aggregation for analysis

Biomedical ontologies

(A quick look at) The Past

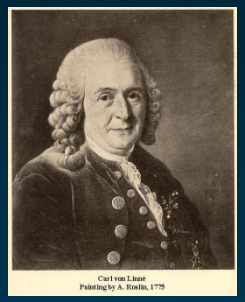
To support a theory of diseases

- Hippocrates
  - Dismisses superstition
  - Four humors
    - Blood
    - Phlegm
    - Yellow bile
    - Black bile
- Thomas Sydenham (1624-1689)
  - *Medical observations on the history and cure of acute diseases* (1676)



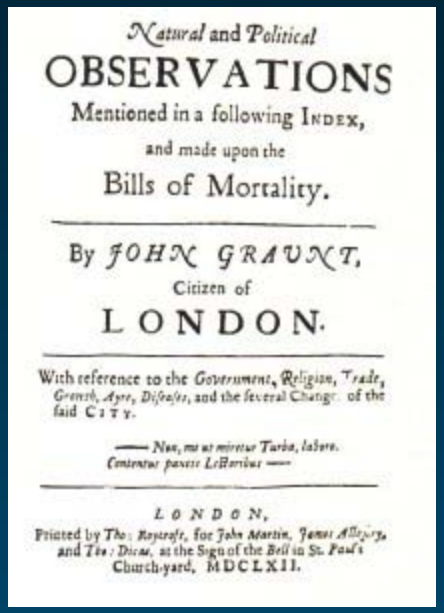
To classify diseases (and plants)

- Carolus Linnaeus (1707-1778)
  - *Genera Plantarum* (1737)
  - *Genera Morborum* (1763)
- François Boissier de La Croix a.k.a. F. B. de Sauvages (1706-1767)
  - *Methodus Foliorum* (1751)
  - *Nosologia Methodica* (1763/68)
- William Cullen (1710-1790)
  - *Synopsis Nosologiae Methodicae* (1785)



To support epidemiology

- John Graunt (1620-1674)
  - Analyzes the vital statistics of the citizens of London
- William Farr (1807-1883)
  - Medical statistician
  - Improves Cullen's classification
  - Contributes to creating ICD
- Jacques Berthillon (1851-1922)
  - Chief of the statistical services (Paris)
  - Classification of causes of death (161 rubrics)


Natural and Political
OBSERVATIONS
Mentioned in a following INDEX,
and made upon the
Bills of Mortality.

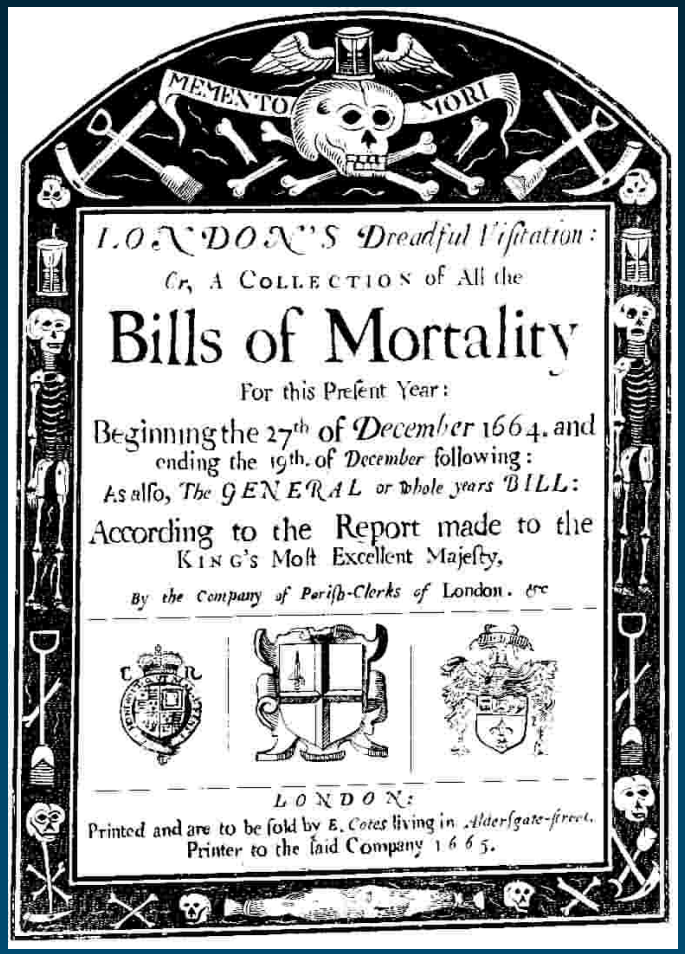
By JOHN GRAUNT,
Citizen of
LONDON.

With reference to the Government, Religion, Trade, Growth, Ayre, Diseases, and the several Changes of the said CITY.

—— Non, me ut miretur Turba, labore.
Contentus paucis Lectoribus ——

LONDON,
Printed by Tho: Roycroft, for John Martin, James Allestry, and Tho: Dicas, at the Sign of the Bell in St. Paul's Church-yard, MDCLXII.

London Bills of Mortality



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 Stillborne	617	Executed	21	Palsie	50
Aged	1545	Flux and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in Streets, fields, &c.	2	Planet	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	19
Bedrid	12	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	Rising of the Lights	397
Burnt and Scalded	8	Hanged & made away themselves	7	Rupture	34
Calenture	3	Headmould shot & Mouldfalleen	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	Sores, Ulcers, broken and bruised	
Childbed	625	Killed by severall accidents	46	Limbs	82
Christomes and Infants	1258	Kings Evill	86	Spleen	14
Cold and Cough	63	Leprosie	2	Spotted Feaver and Purples	1929
Collick and Winde	134	Lethargy	14	Stopping of the Stomack	334
Consumption and Tisick	48[illegible]8	Livergrown	2[illegible]	Stone and Strangury	58
Convulsion and Mother	2056	Megrom and Headach	12	Suddat	[illegible]
Distracted	5	Measles	7	Teeth and Worms	2614
Dropsie and Tympany	1478	Murthered and Shot	9	Vomiting	54
Drowned	53	Overlaid & Starved	45	Vrine	7

Christned: Males	5114	Buried: Males	48568	Of the Plague	68596
Christned: Females	4853	Buried: Females	48737		
Christned: In all	9967	Buried: In all	97305		

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

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.

Biomedical ontologies

The Present

Many biomedical ontologies

- About 200 biomedical ontologies available in various repositories
- Over 2M biomedical concepts
- Hundreds of millions of relations among them
- Limited interoperability
- Quality assurance issues

Many biomedical ontologies

- General vocabularies
  - anatomy (FMA, Neuronames)
  - drugs (RxNorm, First DataBank, Micromedex)
  - medical devices (UMD, SPN)
- Several perspectives
  - clinical terms (SNOMED CT)
  - information sciences (MeSH, CRISP)
  - administrative terminologies (ICD-9-CM, CPT-4)
  - data exchange terminologies (HL7, LOINC)

Many biomedical ontologies (cont'd)

- Specialized vocabularies
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - oncology (PDQ)
  - psychiatry (DSM, APA)
  - adverse reactions (MedDRA, WHO ART)
  - primary care (ICPC)
- Terminology of knowledge bases (OMIM, QMR)

Too many biomedical ontologies?

- Examples of exotic or obsolete ontologies in biomedical ontology repositories
- Governance issues
  - e.g., Ontology developed by a doctoral student

Uses of biomedical ontologies

- Knowledge management
  - Annotating data and resources
  - Accessing biomedical information
  - Mapping across biomedical ontologies
- Data integration, exchange and semantic interoperability
- Decision support
  - Data selection and aggregation
  - Decision support
  - Natural language processing (NLP) applications
  - Knowledge discovery

[Bodenreider, YBMI 2008]

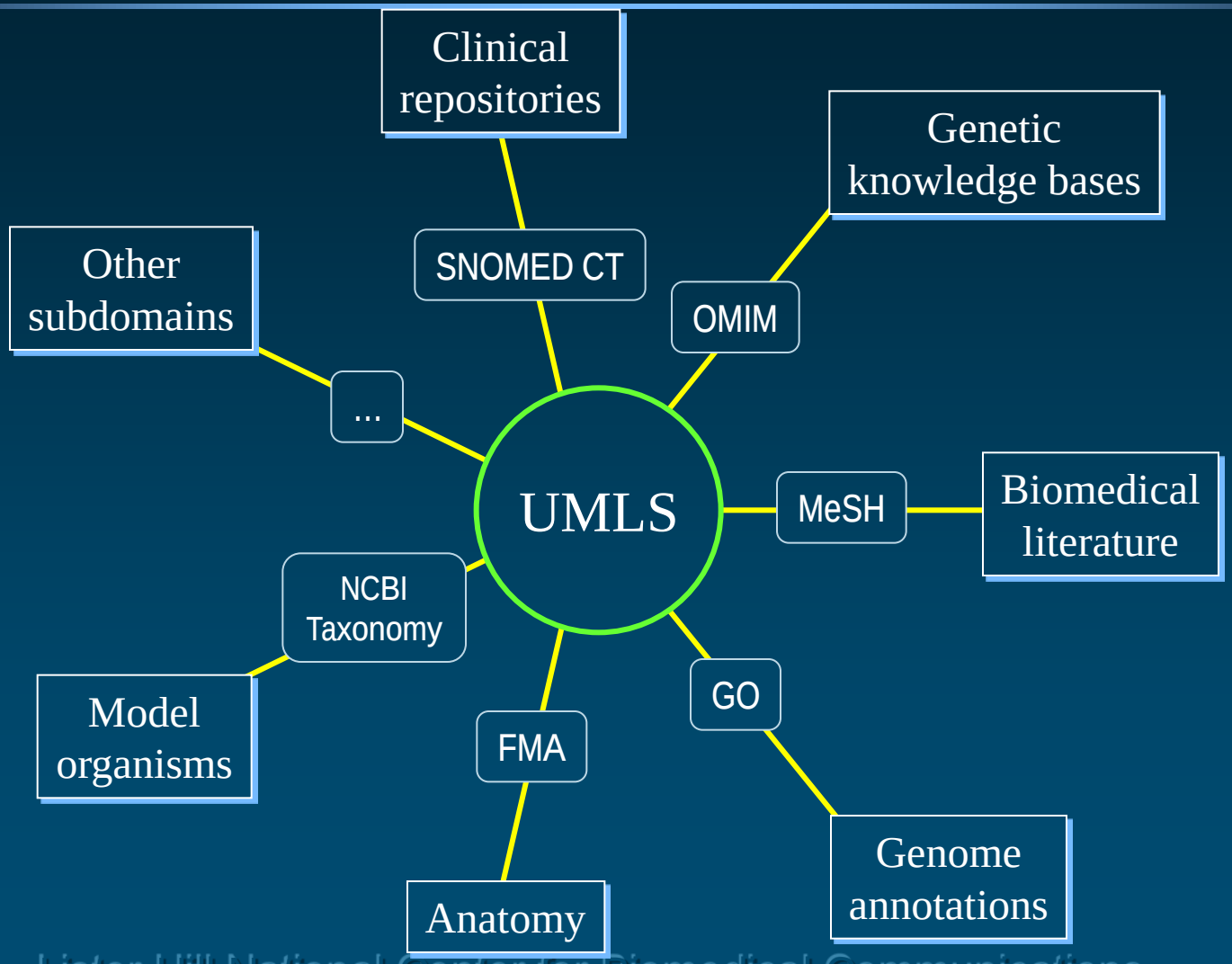
Development

- Still mostly uncoordinated
  - "Cottage industry"
  - Issues
    - Redundancy
    - Lack of consistence
    - Need for mapping
  - Exception: OBO Foundry
- Knowledge representation technology
  - Move towards description logics (e.g., OWL)
  - e.g., SNOMED CT [+ OBO ontologies]

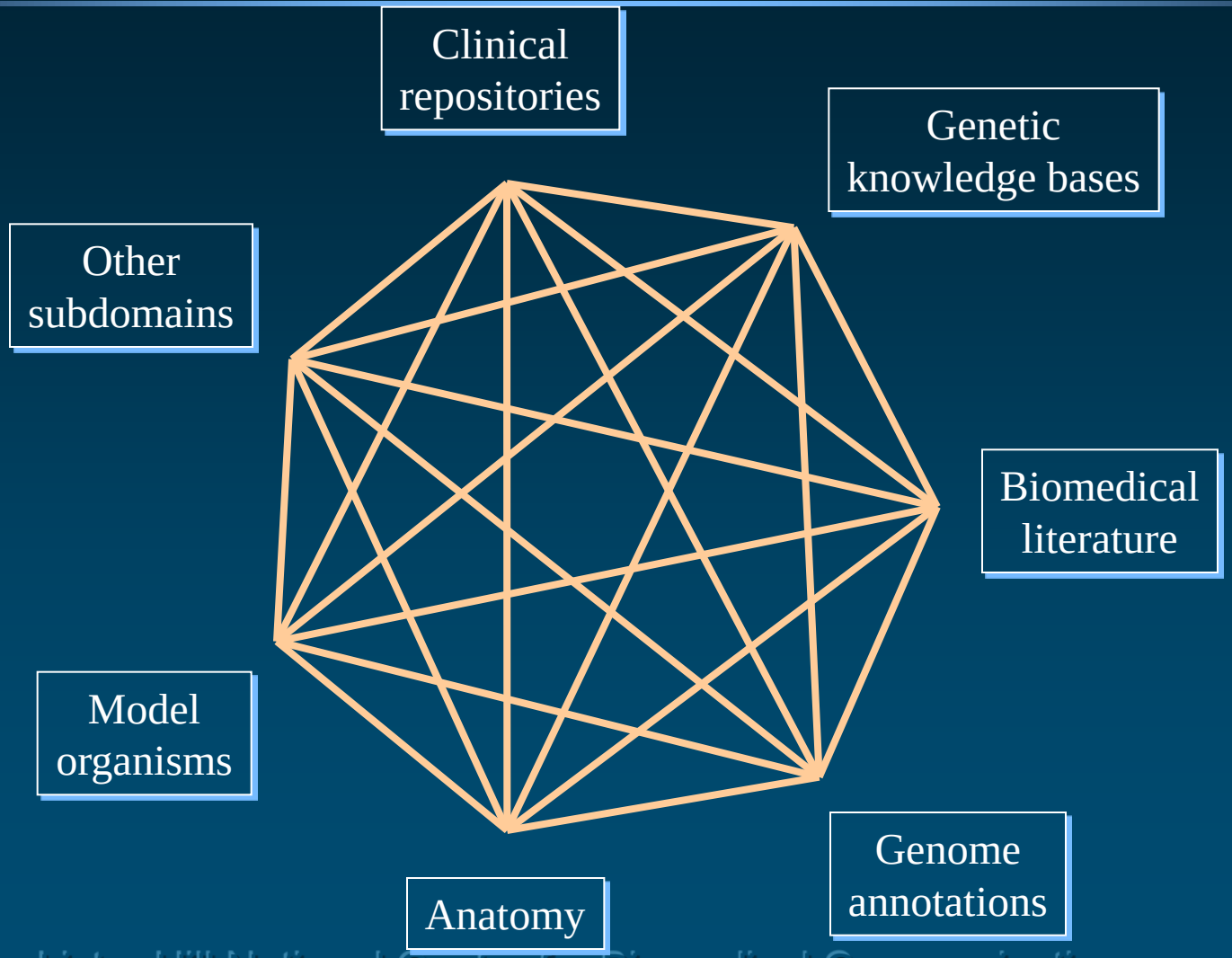
Loose integration

- Pairwise mappings
  - Unidirectional
  - Specific to a given purpose
  - Costly to create and maintain
- Integration through a reference
  - “Interlingua”
  - Identify which terms from different ontologies name the same entities and link them together
  - e.g., Unified Medical Language System (UMLS) Metathesaurus

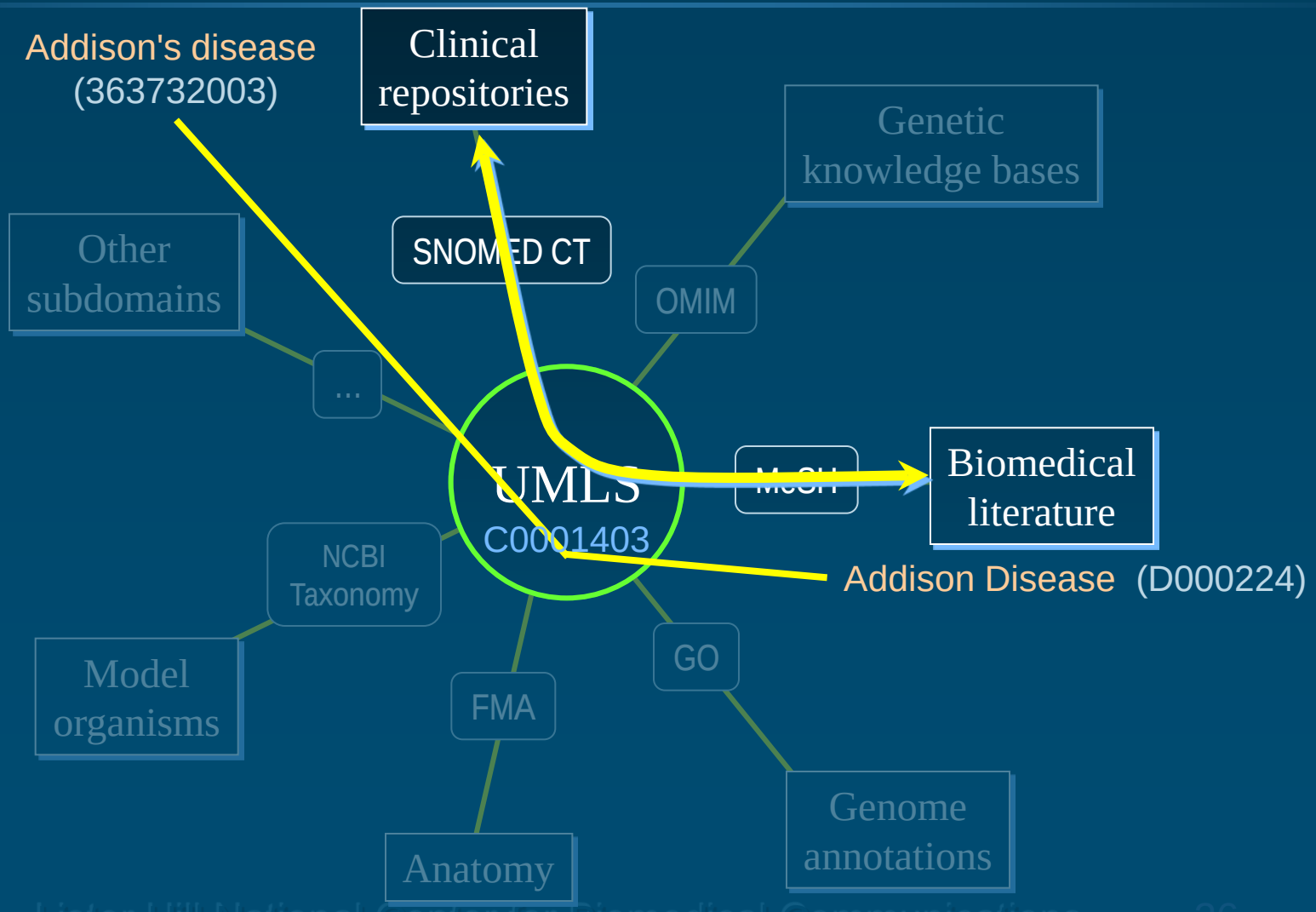
Integrating subdomains



Integrating subdomains



Ontology integration through the UMLS

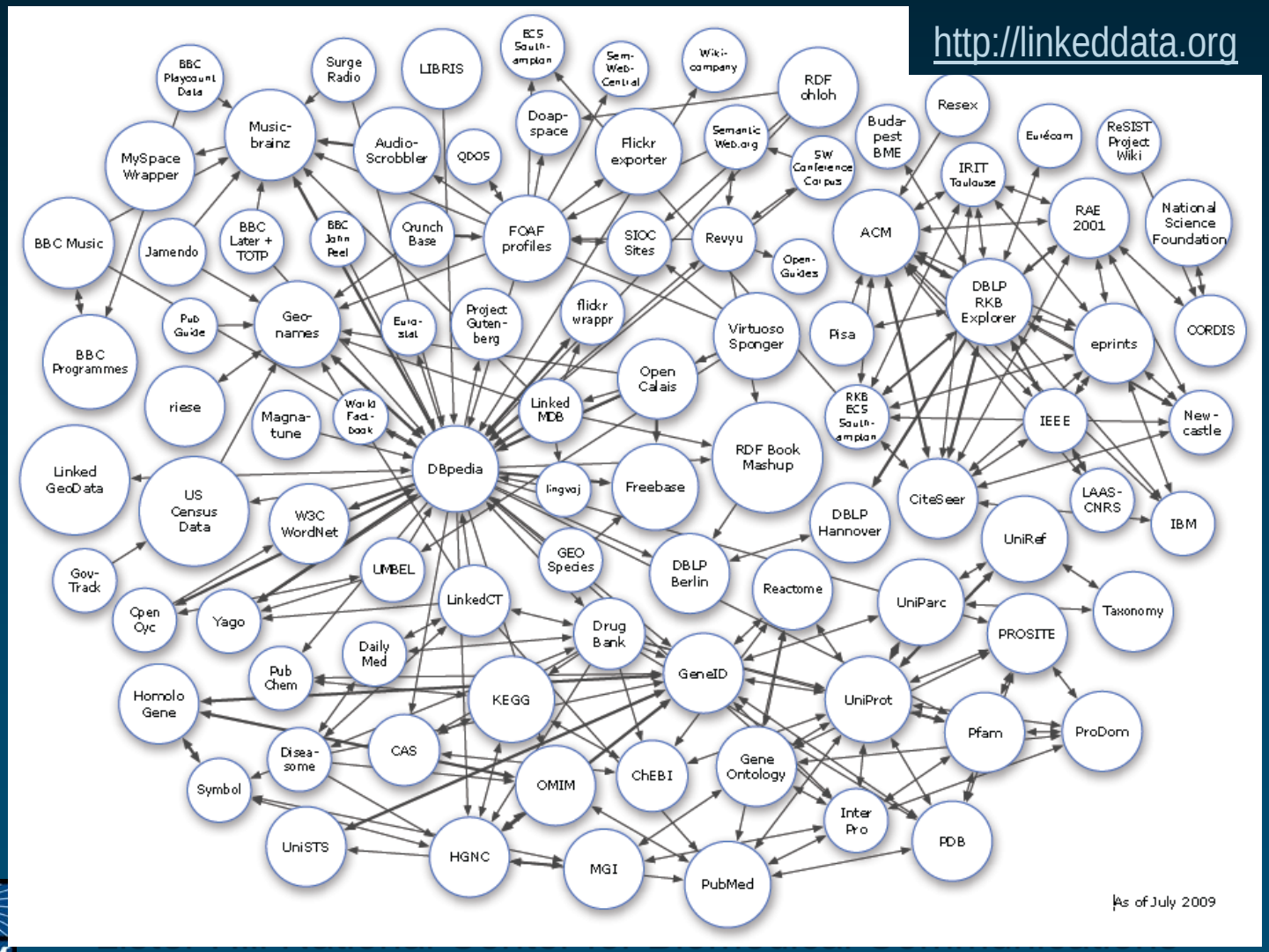


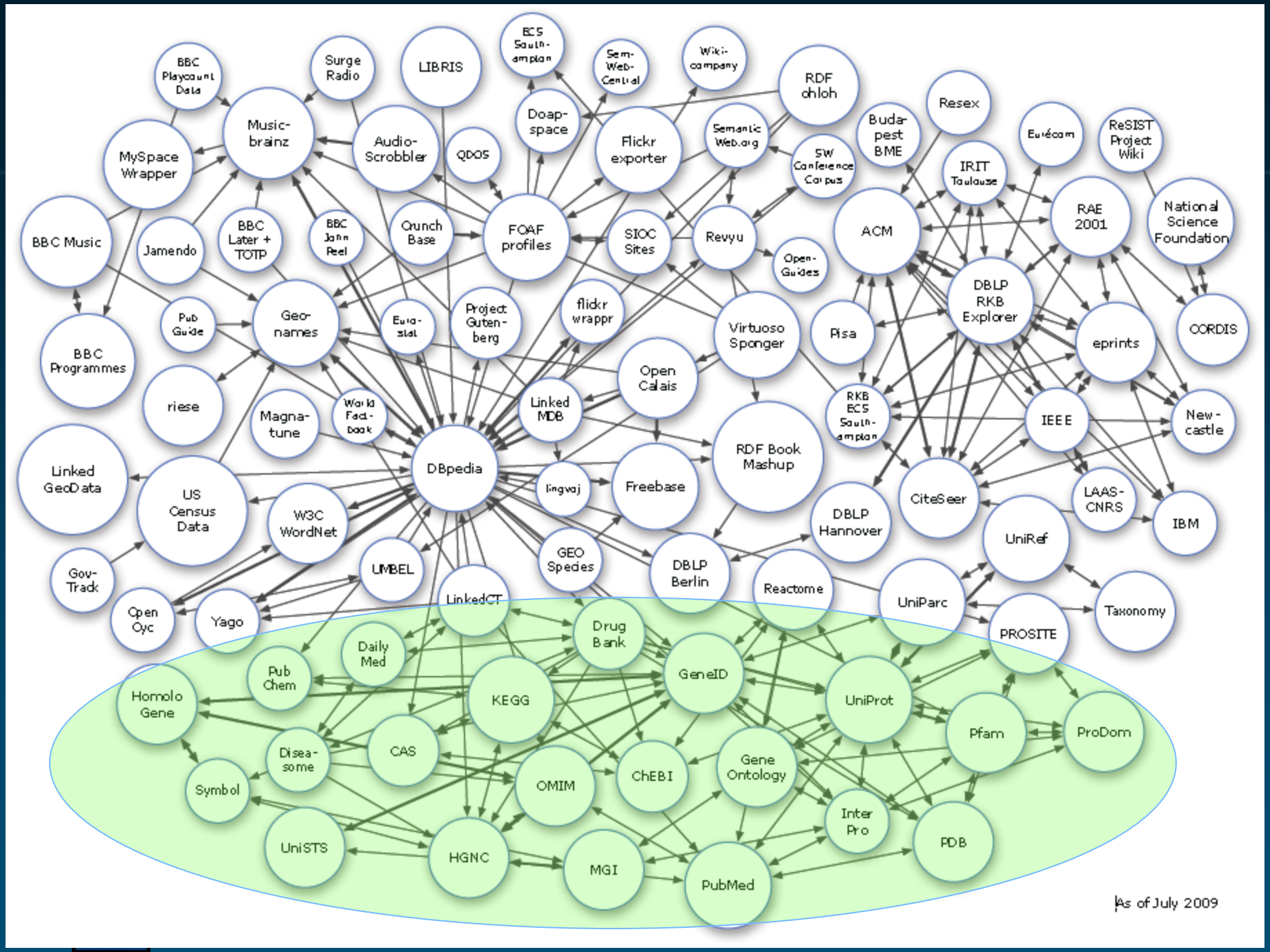
(Integrated) concept repositories

- Unified Medical Language System
  http://umlsks.nlm.nih.gov
- NCBO's BioPortal
  http://www.bioontology.org/tools/portal/bioportal.html
- caDSR
  http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr
- Open Biomedical Ontologies (OBO)
  http://obofoundry.org/

Ontology integration supports data integration

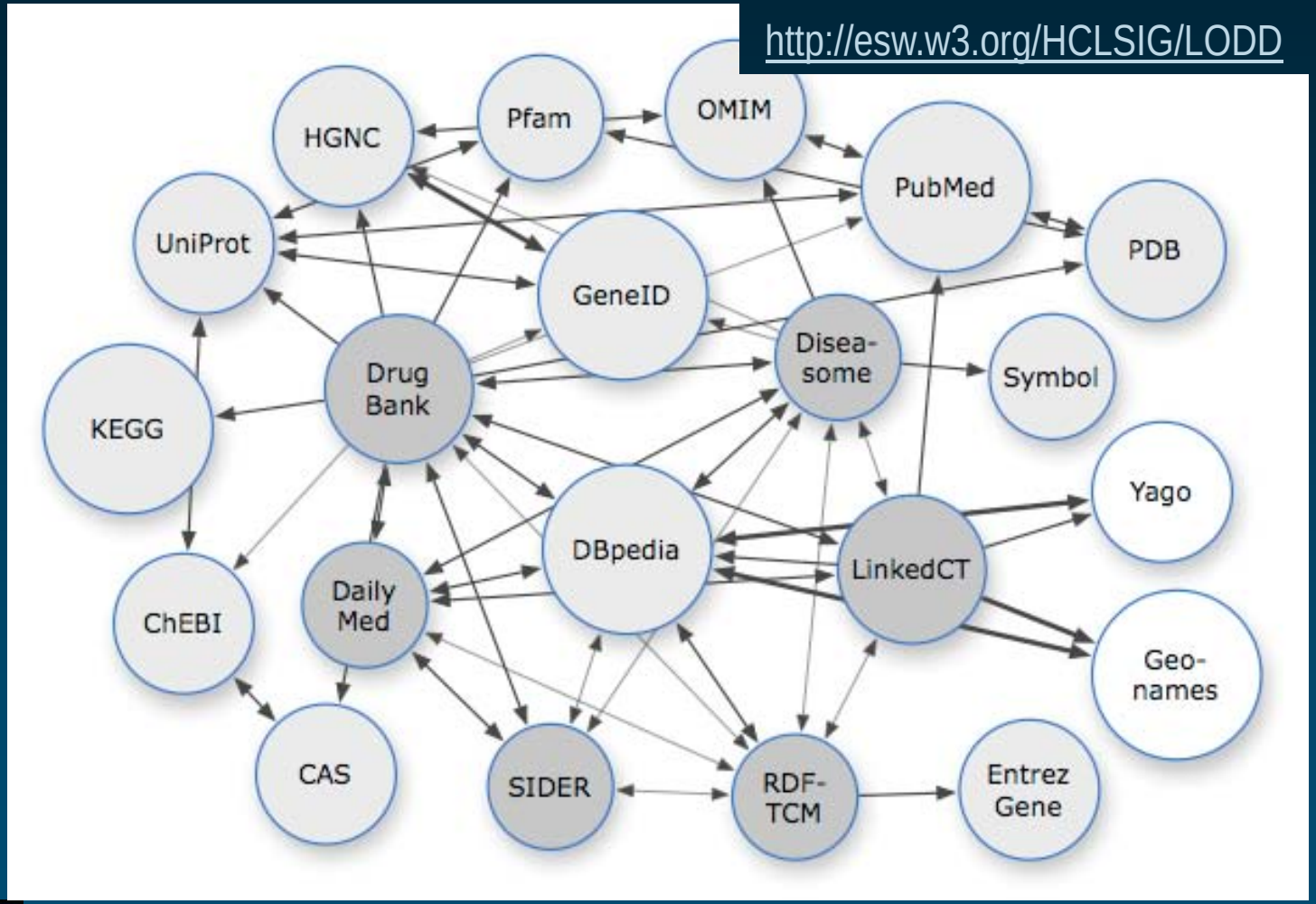
http://linkeddata.org




BBC Playcount Data
Surge Radio
LIBRIS
ECS South-ampton
Sem-Web-Central
Wiki-company
RDF ohloh
Resex
Budapest BME
Eurécom
ReSIST Project Wiki
MySpace Wrapper
Music-brainz
Audio-Scrobbler
QDOS
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
IRIT Toulouse
BBC Music
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
RAE 2001
National Science Foundation
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
DBLP RKB Explorer
eprints
CORDIS
BBC Programmes
riese
Magna-tune
World Fact-book
Linked MDB
Open Calais
RKB ECS South-ampton
IEEE
New-castle
Linked GeoData
US Census Data
DBpedia
lingvoj
Freebase
RDF Book Mashup
CiteSeer
LAAS-CNRS
IBM
W3C WordNet
DBLP Hannover
UniRef
Gov-Track
UMBEL
GEO Species
DBLP Berlin
Reactome
UniParc
Taxonomy
Open Cyc
Yago
LinkedCT
Drug Bank
PROSITE
Daily Med
Pub Chem
GeneID
Homolo Gene
KEGG
UniProt
Pfam
ProDom
Disea-some
CAS
Gene Ontology
Symbol
OMIM
ChEBI
Inter Pro
UniSTS
HGNC
MGI
PDB
PubMed
As of July 2009

Linked Open Drug Data

http://esw.w3.org/HCLSIG/LODD



Linked data

- Semantic Web
- Resources available in RDF
  - Unique, unambiguous identifiers for entities
  - Explicit relations among entities
- Links across resources (federation)
  - Enabled by
    - Shared identifiers across resources
    - Global identifiers, resolvable on the web

Biomedical ontologies

The Future

Harmonization

- Collaboration among ontology developers
  - Prospectively
    - OBO Foundry model
      - Avoid redundancy
      - Foster collaboration
  - Retrospectively
    - SNOMED CT model
      - Seek agreement with other ontologies for specialized content (e.g., LOINC for observables)
      - Serve as an ontology backbone for classifications (e.g., ICD11)

Harmonization Benefits

- Fewer pairwise mapings
  - Not needed for concepts of the same level of granularity
  - Computable automatically for finer-grained concepts
- Increased interoperability
  - Among ontologies
  - Among datasets annotated to these ontologies
  - Among applications using these ontologies

Quality of biomedical ontologies

- Quality assurance in ontologies is still imperfectly defined
  - Difficult to define outside a use case or application
- Several approaches to evaluating quality
  - Collaboratively, by users (Web 2.0 approach)
    - Marginal notes enabled by BioPortal
  - Centrally, by experts
    - OBO Foundry approach
- Related issues
  - Quality of ontology integration (mappings)

Discoverability

- No universal repositories for biomedical datasets
  - Some datasets made available through portals (NCBI, EBI, NCBO)
- Ontology repositories
  - UMLS: 153 source vocabularies (biased towards healthcare applications)
  - NCBO BioPortal: 195 ontologies (biased towards biological applications)
  - Limited overlap between the two repositories
- Need for discovery services
  - Metadata for ontologies and biomedical datasets

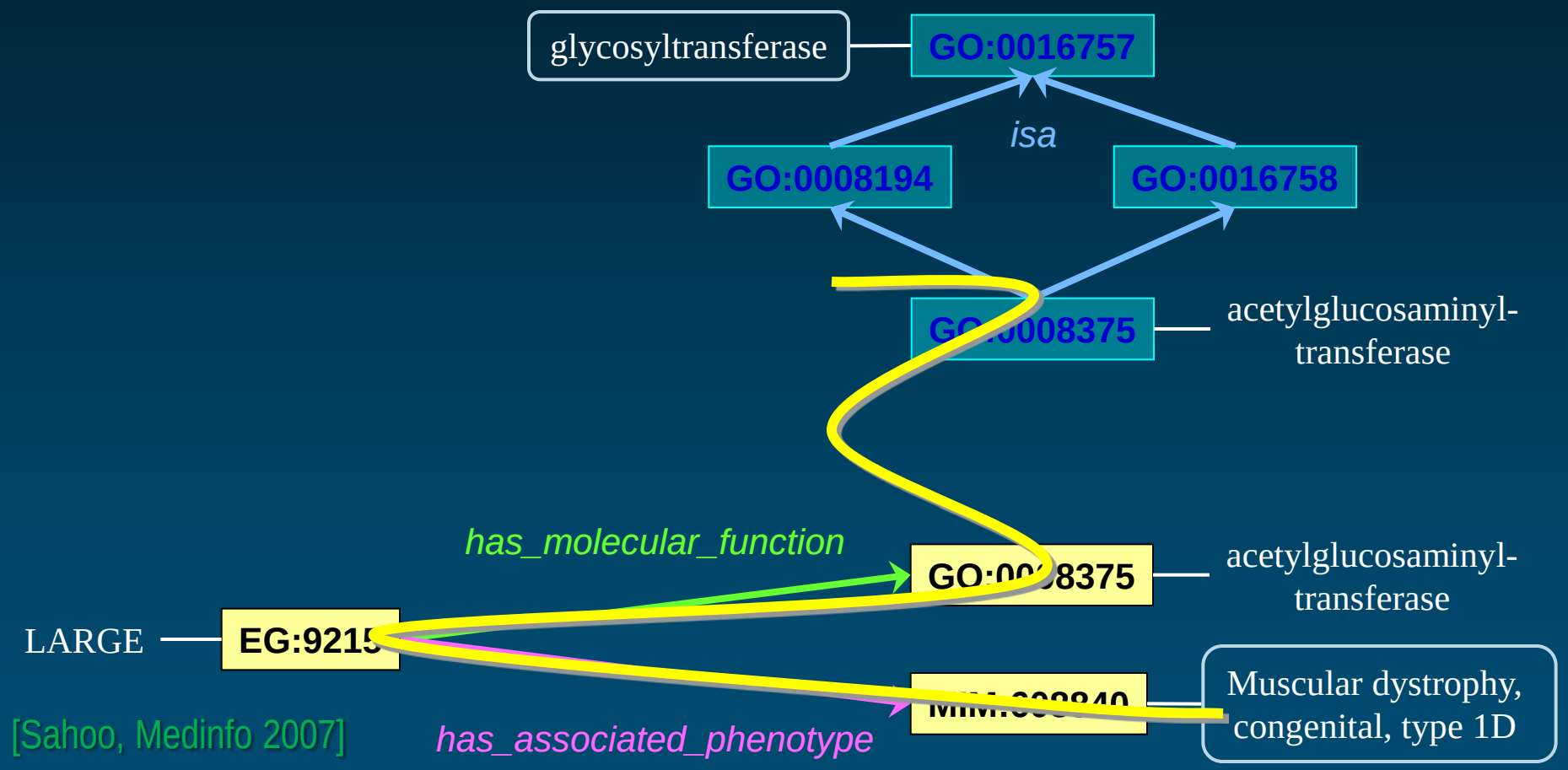
Common upper-level ontologies

- Formalize high-level ontological distinctions
  - Occurrents/continuants
  - Dependent/independent continuants
- Can be shared by multiple domain ontologies
- Make ontologies easier to integrate
  - Fewer essential differences in the organization

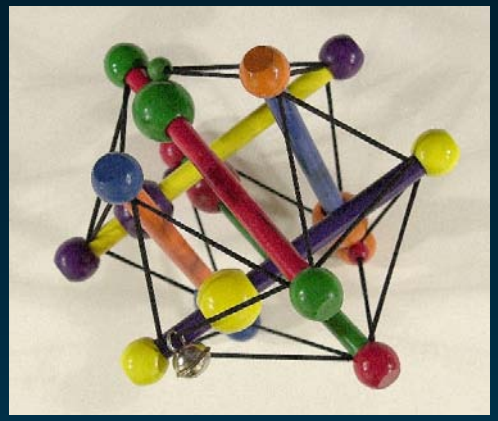
Reasoning with ontologies

- Description logic (e.g., OWL) reasoners available, but few biomedical ontologies can fully take advantage of them
  - Limited expressiveness of the ontologies
  - Limited performance of the reasoners
- Subsumption reasoning
  - Useful for data aggregation
- Beyond subsumption reasoning
  - Rule-base systems (e.g., for clinical decision support)
  - Hypothesis generation and knowledge discovery

From *glycosyltransferase* to *congenital muscular dystrophy*



[Sahoo, Medinfo 2007]



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