

Public Health Informatics Fellowship Program

November 21, 2008

Ontologies and data integration in biomedicine

What is in there for Public Health?



Olivier Bodenreider

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

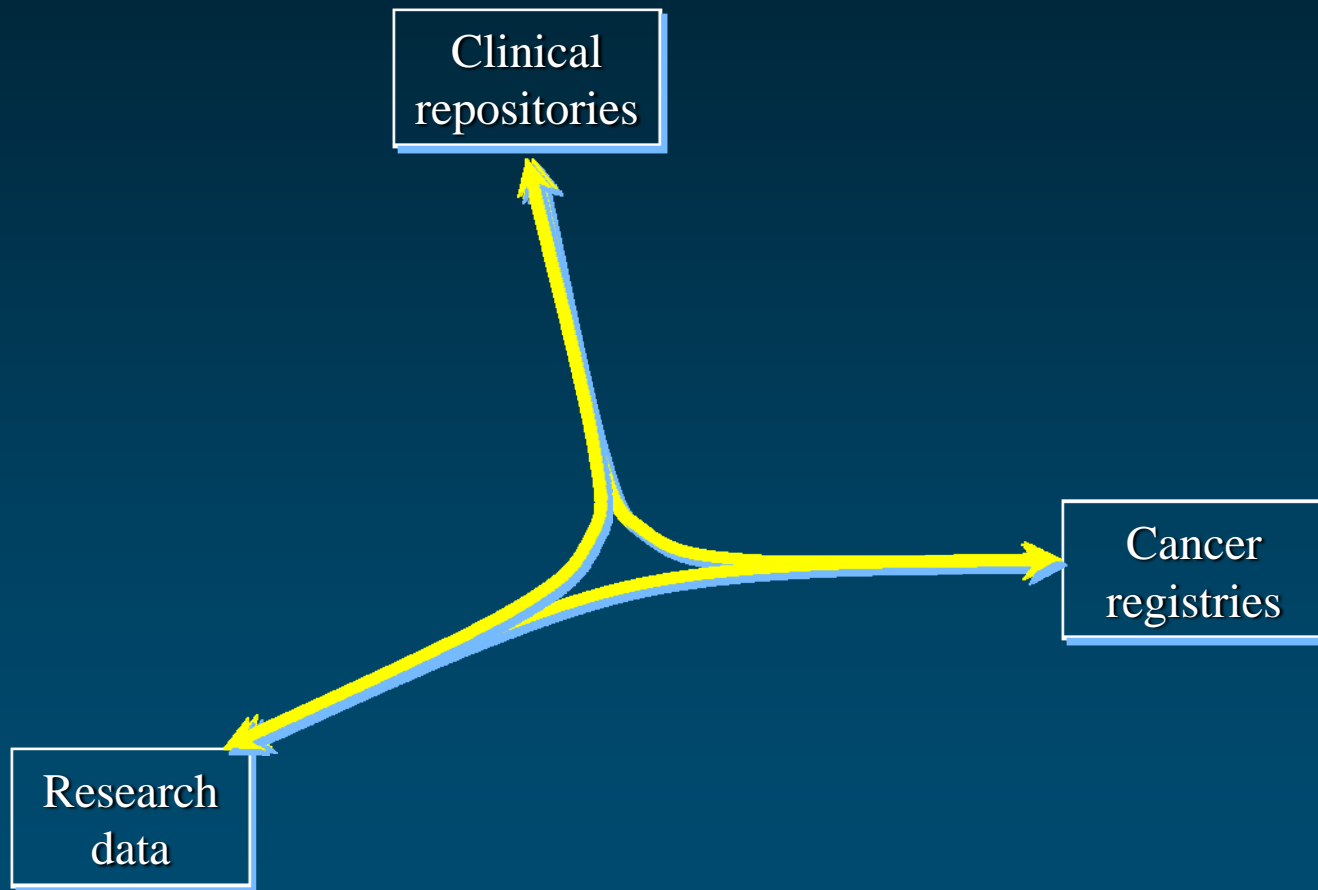
Why integrate data?

Motivation

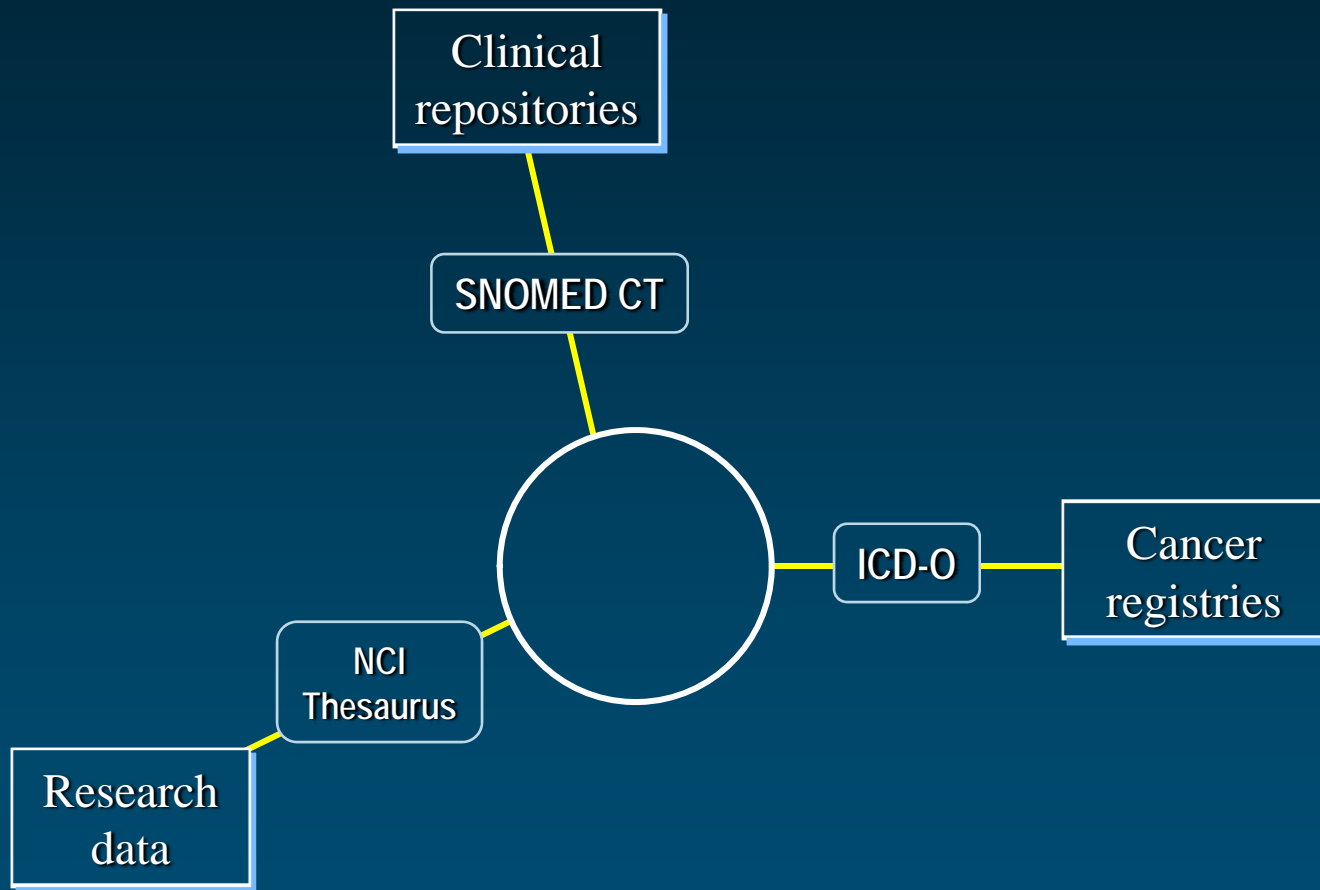
◆ Multiple sources of data

- Each focusing on one aspect
 - Annotated with specific vocabularies
 - Specific format
- Need to be integrated to infer new facts
 - Translational medicine (“Bench to bedside”)
 - Biosurveillance
 - Translation into policy (Public health)

Example Oncology



Integrating subdomains



Example Oncology

◆ Multiple data repositories

- International Classification of Diseases-Oncology (ICD-O-3)
 - Cancer registries
 - Epidemiology, Public health
- SNOMED CT
 - Patient records
 - Clinical care
- NCI Thesaurus
 - Annotation of research data



SNOMED CT



CliniClue 2006: SNOMED CT(International 0707Int[Release]) [Registered user: olivier@nlm.nih.gov]

File Edit Subsets Restrict Language Layout Tools Help

Concept Id 399490008 **Adenocarcinoma of prostate**

Description Id 1778899017 clinical finding

Words - any order

Find prostate adenocarcinoma

P adenocarcinoma of prostate

Hierarchy Subtype hierarchy

- C 254900004 carcinoma of prostate
- C 423748001 adenocarcinoma of pelvis
 - E 399490008 adenocarcinoma of prostate
 - C 278060005 endometrioid carcinoma of prostate

adenocarcinoma of prostate - Definition

Concept Status: **Current**

Descriptions

- F adenocarcinoma of prostate (disorder)
- P adenocarcinoma of prostate

Definition: Fully defined by ...

- is a
 - D carcinoma of prostate
 - D adenocarcinoma of pelvis
- Group
 - associated morphology
 - D malignant adenomatous neoplasm - category
 - finding site
 - D prostatic structure
- Group
 - associated morphology
 - D carcinoma
 - finding site
 - D prostatic structure

Qualifiers

- episodicity
 - P episodicities

Codes

- Original SnomedId : D7-F046E
- Read Code (Ctv3Id) : XUYqi

<http://www.clinical-info.co.uk/>

NCI Thesaurus



Concept Details

URI: http://ncitterms.nci.nih.gov:80/NCIBrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&code=C2919
Version: June 2007 (07.06d)

Prostate Adenocarcinoma

Identifiers:

name	Prostate_Adenocarcinoma
code	C2919

Relationships to other concepts:

Disease_Has_Finding	Invasive Lesion
Disease_Has_Abnormal_Cell	Adenocarcinoma_Cell
Disease_Has_Normal_Tissue_Origin	Prostatic Epithelium
Disease_May_Have_Finding	Serum Prostate Specific Antigen Increased
Disease_Has_Finding	Carcinomatous Component Present
Disease_Excludes_Abnormal_Cell	Neoplastic Smooth Muscle Cell
Disease_Excludes_Abnormal_Cell	Malignant Squamous Cell
Disease_Has_Primary_Anatomic_Site	Prostate Gland
Disease_Has_Associated_Anatomic_Site	Male Reproductive System
Disease_Excludes_Abnormal_Cell	Malignant Stromal Cell
Disease_Has_Associated_Anatomic_Site	Prostate Gland
Disease_Has_Normal_Cell_Origin	Epithelial Cell

Information about this concept:

DEFINITION

Synonym with source data

Synonym with source data

Synonym with source data

Preferred_Name

Semantic_Type

Synonym

Synonym

Synonym

Unified Medical Language System Concept Identifier

Superconcepts:

Adenocarcinoma
Common Carcinoma
Invasive Prostate Carcinoma

Subconcepts:

Acinar Prostate Adenocarcinoma
Metastatic Prostatic Adenocarcinoma
Moderately Differentiated Prostate Adenocarcinoma
Poorly Differentiated Prostate Adenocarcinoma
Prostate Adenocarcinoma with Focal Neuroendocrine Differentiation
Prostate Ductal Adenocarcinoma
Stage III Prostate Adenocarcinoma
Stage II Prostate Adenocarcinoma
Stage I Prostate Adenocarcinoma
Well Differentiated Prostate Adenocarcinoma

Quick Search

Advanced Search

Max Results: 25

prostate adenocarcinoma



ICD-O-3



◆ Morphology

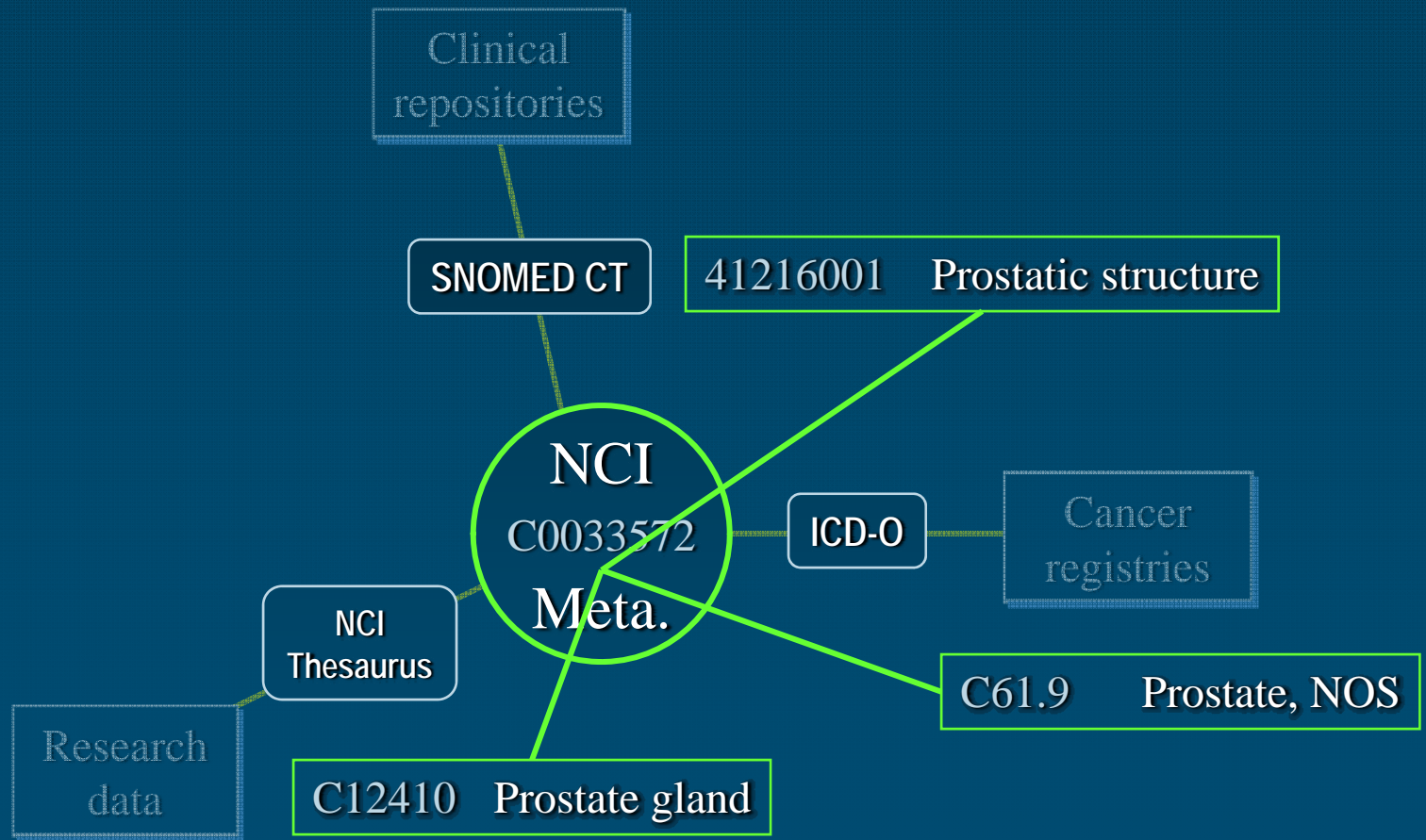
- [...]
- 814-838 Adenomas and adenocarcinomas
 - 8140/3 Adenocarcinoma, NOS

◆ Anatomy

- [...]
- C60-C63 Male genital organs
 - C61 Prostate gland
 - C61.9 Prostate, NOS
Prostate gland

Adenocarcinoma
of prostate

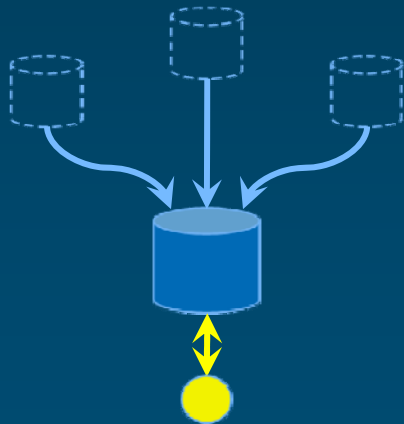
Terminology integration NCI Metathesaurus



Approaches to data integration

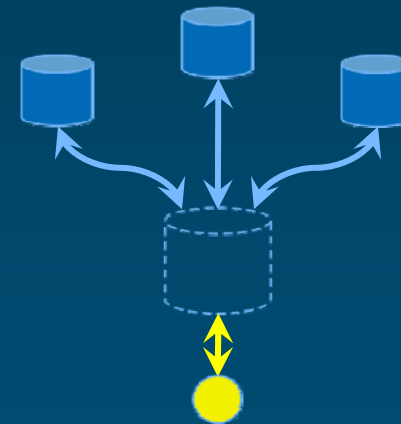
◆ Warehousing

- Sources to be integrated are transformed into a common format and converted to a common vocabulary



◆ Mediation

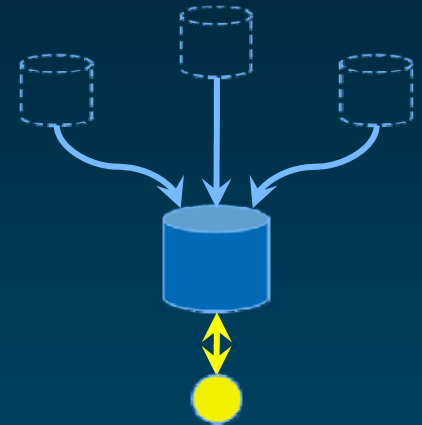
- Local schema (of the sources)
- Global schema (in reference to which the queries are made)



Ontologies and warehousing

◆ Role

- Provide a conceptualization of the domain
 - Help define the schema
 - Information model vs. ontology
- Provide value sets for data elements
- Enable standardization and sharing of data



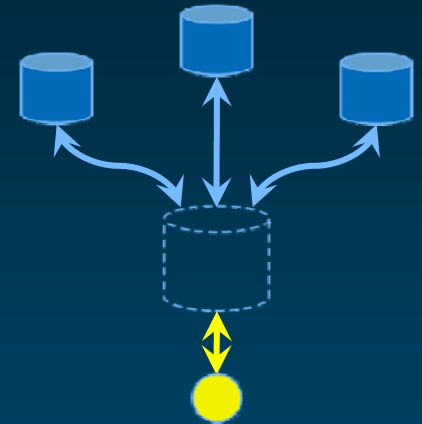
◆ Examples

- BioSense (original design)
- Repositories for translational research (CTSA)
- Clinical information systems

Ontologies and mediation

◆ Role

- Reference for defining the global schema
- Map between local and global schemas

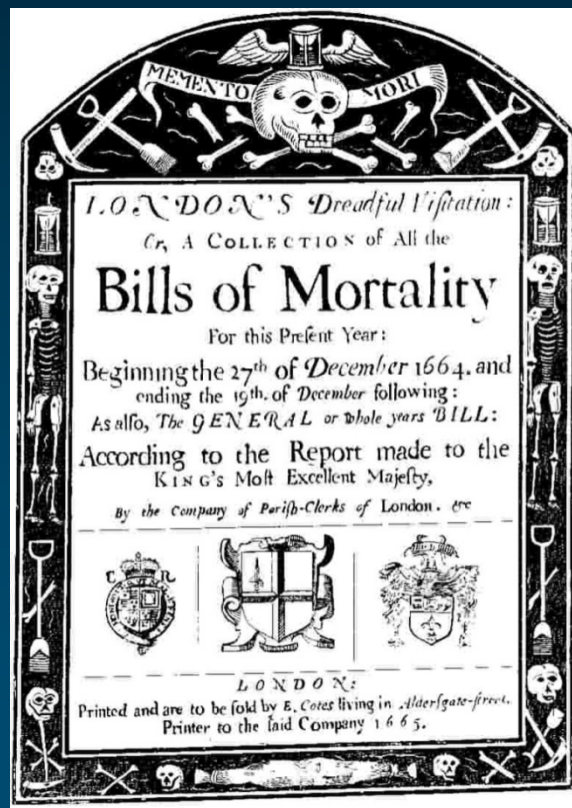


◆ Examples

- BioSense (redesign)
- BioMediator
- OntoFusion

Ontologies Normalization

◆ From the London Bills of Mortality...



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

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

The Diseases and Casualties this year.

A Bortive and Stillborne	519	Executed	31	Pallic	30
Aged	1545	Flox and Small Pox	665	Plague	685
Aque and Peaver	525	Found dead in Streets, fields, &c.	2	Plagues	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	19
Bedrid	12	Frighted	25	Poliozie	4
Blasph	5	Gout and Sciatica	27	Quinsie	35
Bleeding	16	Grief	46	Rickets	157
Bloody Flux, Scouring & Flux	184	Gripping in the Guts	1238	Killing of the Lights	157
Burnt and Scalded	8	Hang'd & made away themselves	7	Rupture	14
Calenture	3	Head-mole & Hor. & Moxle fallen	14	Scurvy	127
Cancer, Gangrene and Fiftula	56	Jaundies	120	Shingles and Swine pox	2
Canker, and Thrush	121	Imposume	227	Sties, Ulcers, broken and boiled	2
Childbed	624	Kild by severall accidents	46	Limbs	82
Cholmes and Infants	1258	Kings Evil	82	Spleen	14
Cold and Cough	62	Leprosie	2	Spotted Fever and Purples	1229
Collick and Winde	134	Lethargy	14	Scouring of the Bowels	314
Consumption and Tiblick	4808	Livergrowne	21	Stone and Strangury	8
Convulsion and Morice	1052	Measles and Headach	12	Sucket	122
Distracted	9	Measles	7	Teeth and Worms	2014
Droove and Turpany	1478	Mothered and Shot	9	Vomiting	34
Drowned	5	Overjaud & Starved	45	Vunn	7
Emiles	5114				
Emiles	4853	Buried	4856		
In all	9967		4837	Of the Plague	685
			2306		

Increased in the Burials in the 130 Parishs and at the Pest-house this year. 7000
Increased of the Plague in the 130 Parishs and at the Pest-house this year. 6850

Ontologies Normalization

- ◆ ...to HITSP vocabulary specifications for interoperability

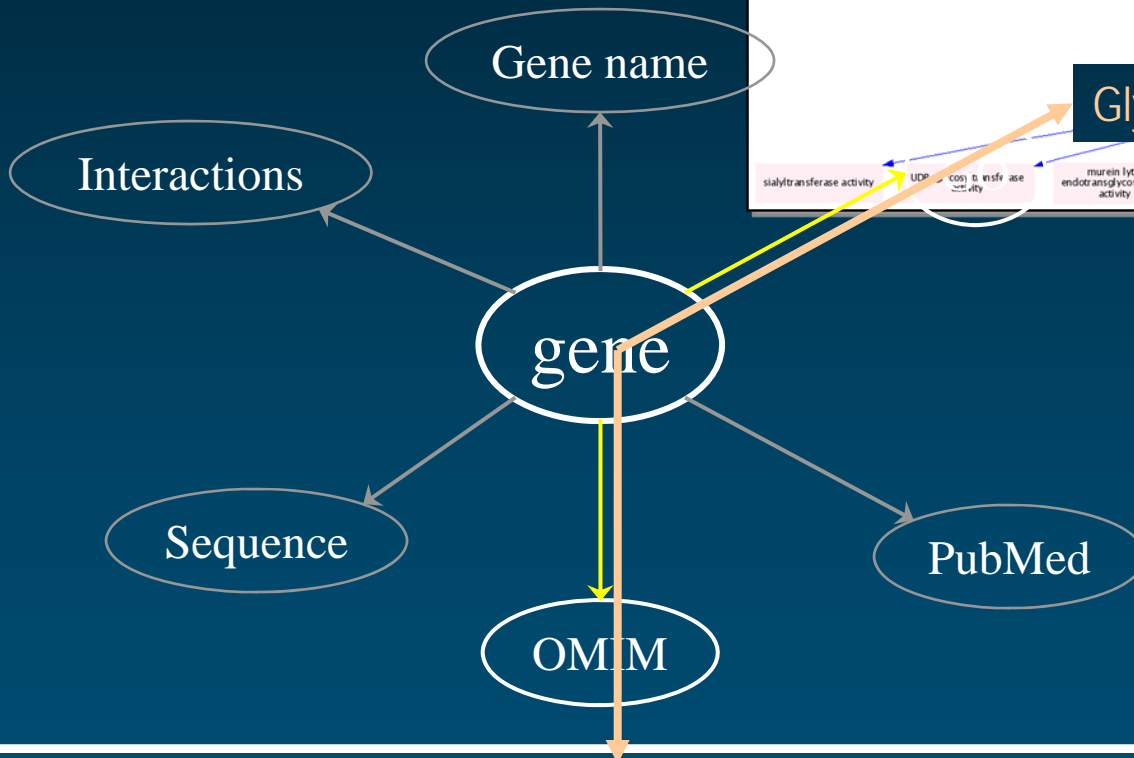
<http://www.hitsp.org/>

Table 2.2.1.3.3-10 Medication Brand Name Vocabulary

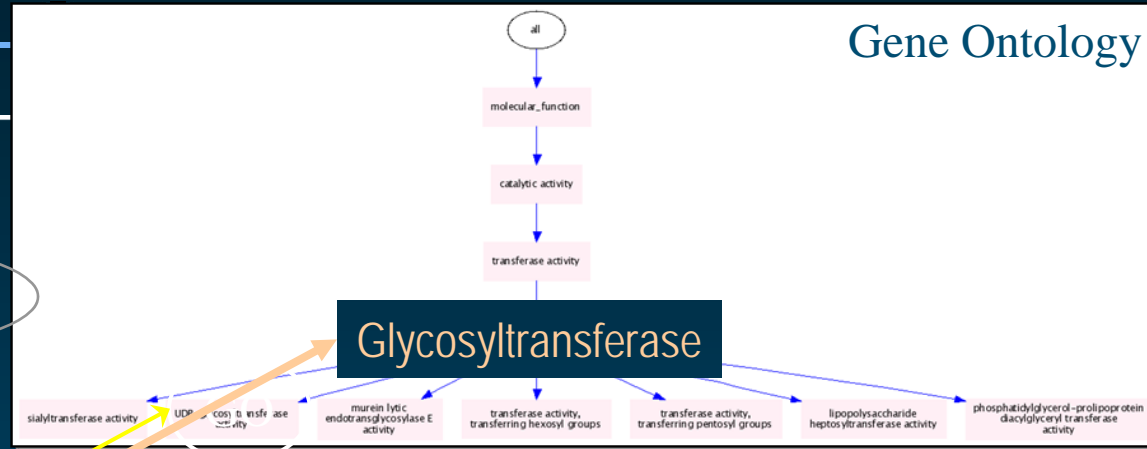
Value Set Name	Medication Brand Name
Vocabulary	Federal Medication Terminologies – RxNorm http://www.cancer.gov/cancertopics/terminologyresources/page4
Source	NLM
Vocabulary OID	2.16.840.1.113883.6.88
Vocabulary Version	
Value Set OID	2.16.840.1.113883.3.88.12.80.16
Value Set Version	
Static/Dynamic	Dynamic
Value Set Members	See http://www.nlm.nih.gov/research/umls/rxnorm/ . Shall contain a value descending from RxNorm.

Ontologies Inference

Entrez Gene



Gene Ontology



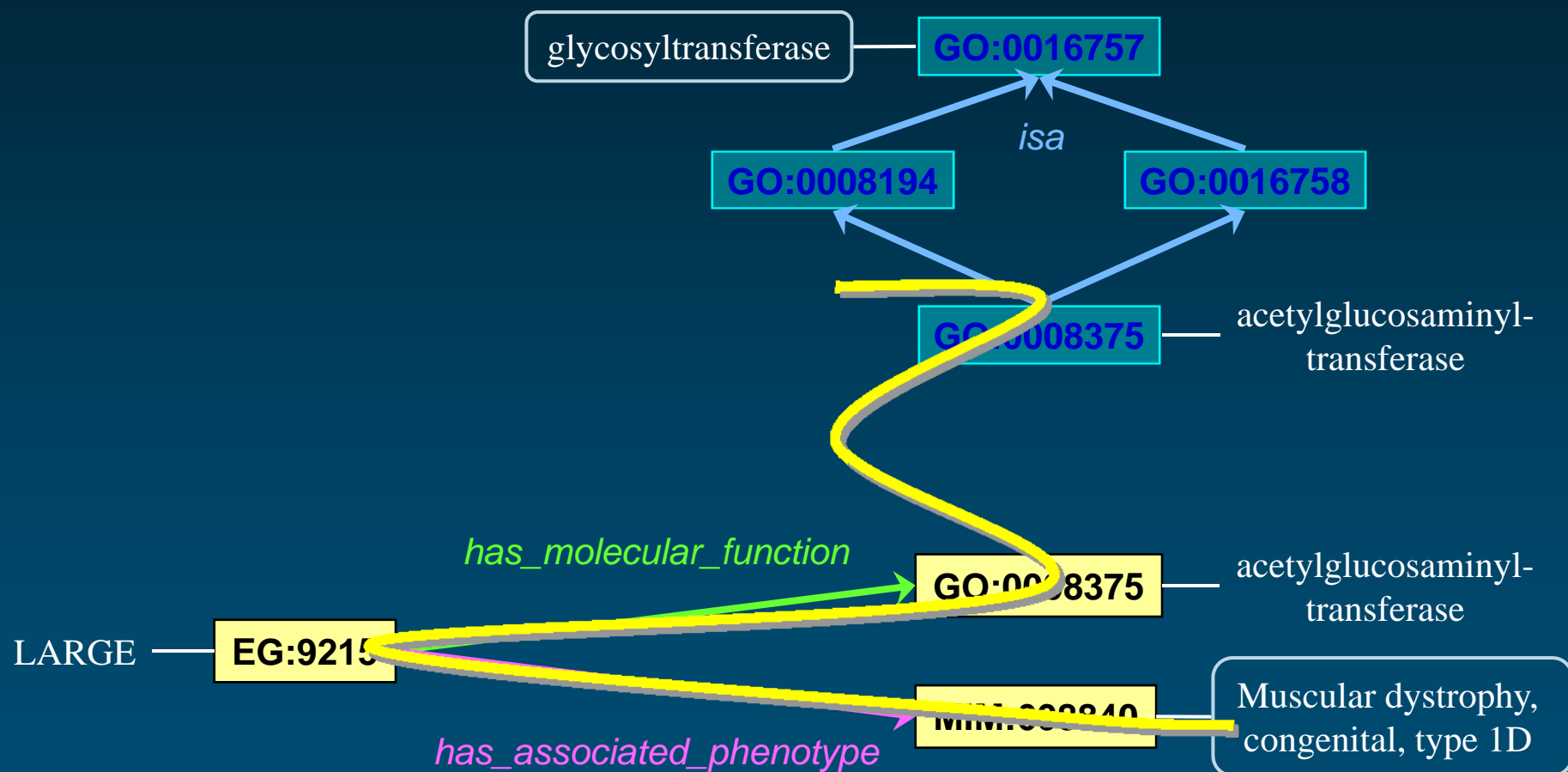
[Sahoo, Medinfo 2007]

Congenital muscular dystrophy



Lister Hill National Center for Biomedical Communications

From *glycosyltransferase* to *congenital muscular dystrophy*





Ontologies, data integration and public health

Case study: BioSense



Results 1 - 10 of about 245

[Service Oriented Architecture: A Practical Approach to Data ...](#)

08/15/06 Linh H. Le, MD, MPH New York State Department of Health SOA and **Data Integration** ... New York State Department of Health Why SOA for **Data Integration** ...

www.cdc.gov/nceh/tracking/tracks06/pdfs/presentation16_le.pdf

[Eliminating Lead](#)

Poisoning through Improved **Data Integration** Wendy Blumenthal, MPH Environmental Public Health Tracking ... and reporting **Data integration** across multiple programs, ...

www.cdc.gov/nceh/tracking/conf04/pdfs/thu/ses2B/w_blumenthal.pdf

[2000 EHDl State Planning Meeting Agenda](#)

Saul Franklin 4:30 - 5:00 Discussion DAY 2 - THURSDAY, NOVEMBER 9, 2000 8:30 - 10:15 Session V: **Data Integration** Activity IH: Integrate ...

www.cdc.gov/ncbddd/ehdi/documents/2000agenda.pdf

[BioSense Redesign](#)

Leverage Existing **Data Integration**: Locals/States/EHIO Data acquisition efforts. Leverage Existing Data: ... Linux. 46. Toolkit Lead. **Data Integration** Maintainer (Lead ...

www.cdc.gov/biosense/files/BioSense_talk_3.19.ppt

[Genomics|Links|Open Source Projects|infrastructure](#)

fields. Adopting controlled vocabulary is a critical step toward **data integration** and interoperability in any information system. ...

www.cdc.gov/genomics/links/open_source_projects/opensource_01.htm

[BioSense - Extramural Projects](#)

Boston University SPH. Improving Syndromic Surveillance by **Data Integration**. Boston University School of Public Health proposed a ...

www.cdc.gov/biosense/extramuralprojects.htm

[Information System Architectures for Syndromic Surveillance](#)

Data-Integration Components. ... These features are implemented in the model systems, but not always with comparable algorithms. Architectures for **Data Integration**. ...

www.cdc.gov/mmwr/preview/mmwrhtml/su5301a37.htm



BioSense Integrating multiple datasets

- ◆ Acute care data (emergency room)
 - Demographic data
 - Clinical data
 - Chief complaints
 - Initial / final diagnosis
 - Laboratory data
- ◆ Pharmacy data (national pharmacies)
- ◆ Laboratory data (national labs)
- ◆ Poison control center data
- ◆ ...



Case study Materials and methods

◆ Materials

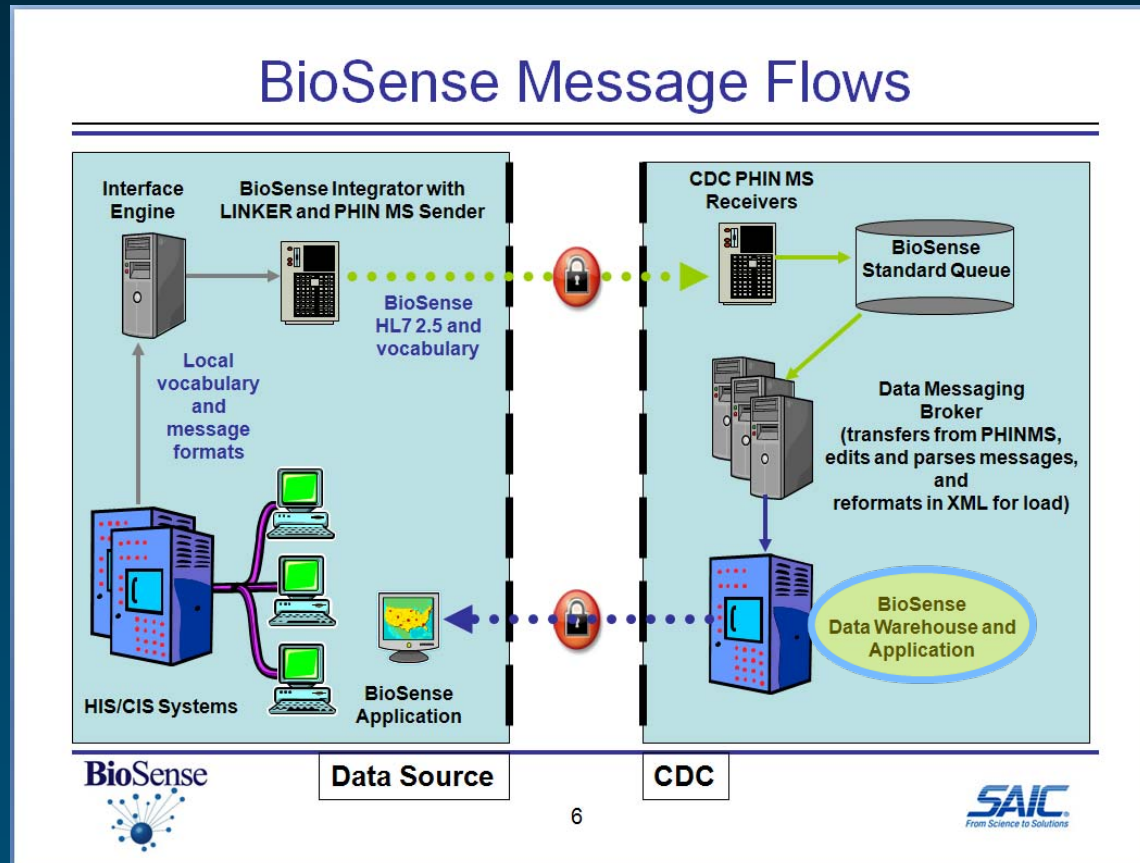
- BioSense publications and presentations

◆ Methods: look for mentions of

- Data integration approaches
- Use of ontologies / terminologies / vocabularies
 - Normalization
 - Inference

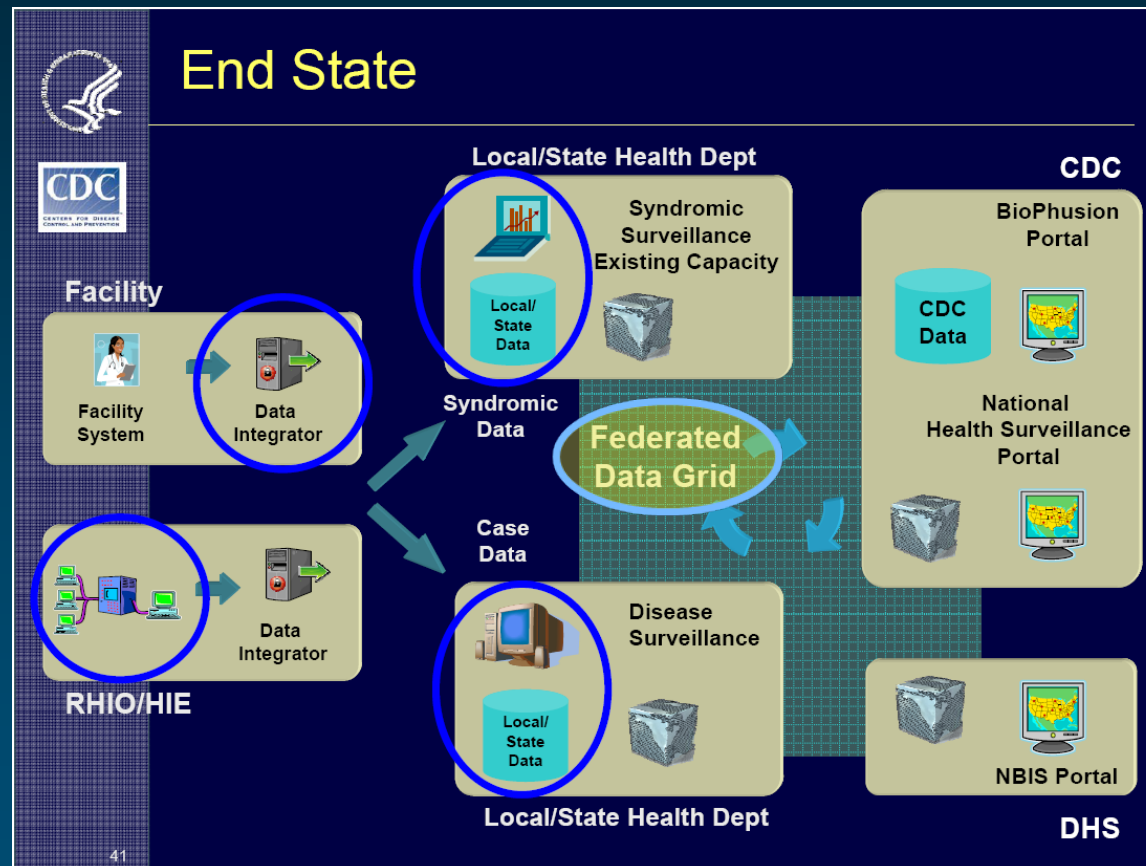
BioSense Approaches to data integration

◆ Initial approach: Warehouse



[Chambers,
PHIN 2008]


◆ Redesign: Mediation (Federation)



[Lenert, ASTHO Roudtable 2008]

BioSense Ontologies

◆ Normalized vocabulary for data exchange



The graphic features a blue header with the word "BioSense" in white. Below it, the text "Frequently Asked Questions" is displayed. To the right is a stylized network diagram with a central blue sphere and radiating lines connecting to smaller blue spheres. The background of the header has a subtle grid pattern.

Is BioSense coordinating with the Office of the National Coordinator for Health Information Technology (ONC) and the American Health Information Community (AHIC) efforts?

Yes, BioSense supports the efforts of the ONC and the Health Information Technology Standards Panel (HITSP) which is a collaborative effort to **harmonize health information interoperability standards, particularly health vocabulary** and messaging standards. AHIC was formed to help advance efforts to reach the President's call for most Americans to have electronic health records within 10 years.

BioSense Ontologies

- ◆ Normalized vocabulary for data exchange
 - Link data elements to value sets from standard vocabularies



Interoperability Specification

IS 02 - HITSP Biosurveillance Interoperability Specification

This Interoperability Specification focuses on a set of constrained standards for implementation of near real-time, nationwide public health event monitoring to support early detection, situational awareness and rapid response management across care delivery, public health and other authorized Government agencies. It prescribes the process or interaction that each primary stakeholder will invoke to capture, discover, anonymize and transmit relevant data.

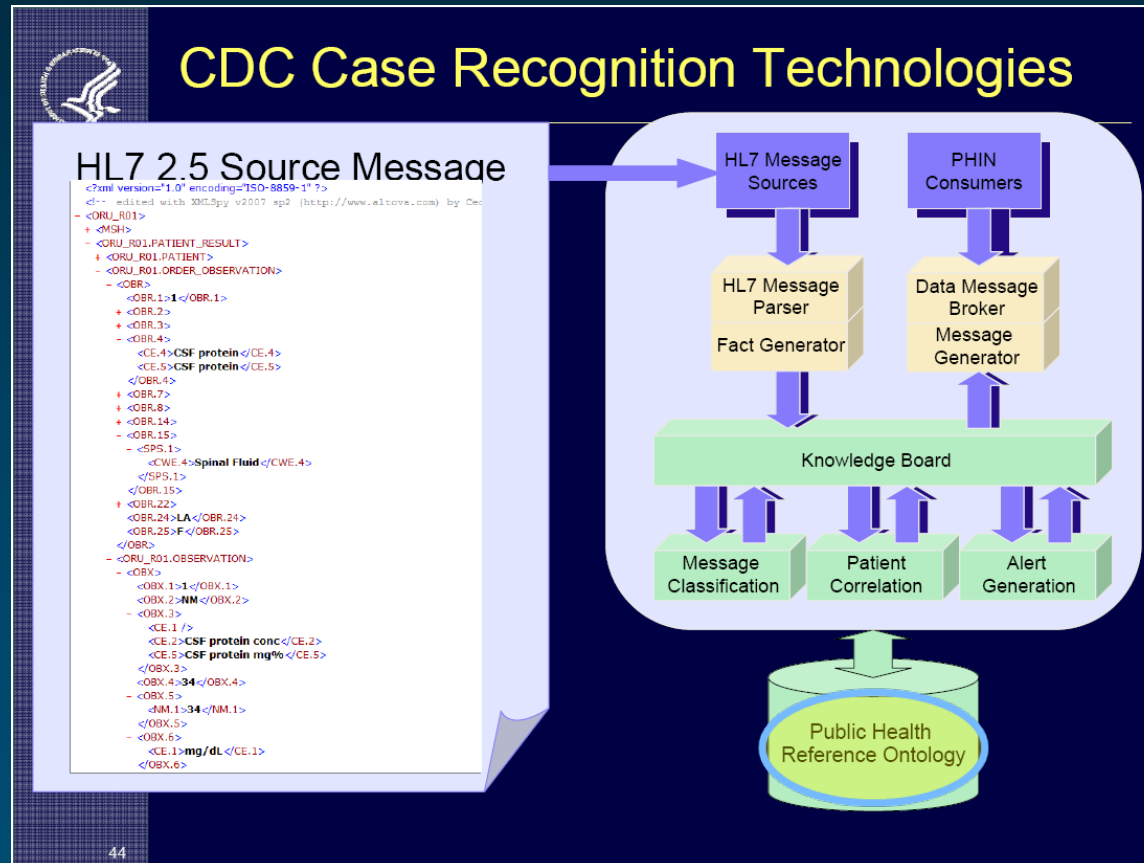
<http://www.hitsp.org/>

[View by Topic](#) ✓
[View by Status](#) ✓
[View Complete Library](#) ✓

Construct	Title / Version	Referenced by	Status	Document Access
IS 02	HITSP Biosurveillance Interoperability Specification Version:2.1	IS 2 V:2.1	Recognized	DOWNLOAD 
IS 02	HITSP Biosurveillance Interoperability Specification Version:3.0	IS 2 V:3.0	Released (Panel Approved)	DOWNLOAD 
IS 02	HITSP Biosurveillance Interoperability Specification Version:3.1	IS 2 V:3.1	Released (Panel Approved)	DOWNLOAD 

BioSense Ontologies

◆ Ontologies for aggregation and inference



[Lenert, ASTHO
Roudtable 2008]

Public Health Reference Ontology

**ph in**
PUBLIC HEALTH INFORMATION NETWORK

Conference 2008

**Public Health Informatics:
Collaboration at the Crossroads**

Start | Browse by Day | Author Index

Ontology Engineering Application of Reasoning Services

Tuesday, August 26, 2008: 3:30 PM
International B

Craig Cunningham, BS, Computer, Science , *OntoReason LLC, Salt Lake City, UT*
Gautam Kesarinath, MS , *NCPHI, Centers for Disease Control and Prevention, Atlanta, GA*

[Slides](#) [Slides](#)

This workshop session will present a detailed discussion based upon the reasoning services that have been developing using the public health reference ontology as the domain knowledgebase. This training workshop will discuss knowledge representation within a reasoning framework including the discussion of rule templates and advanced artificial intelligence techniques. In addition to the basis-reasoning platform, ontology knowledge representations, and reasoning techniques, the presentation will cover the analysis of three ontology-based reasoning services. Message Content Validation Reasoner, Case Classification, and BioSense Message Classification. Also presented in the training workshop will be the knowledge components and services that provide an extensible foundation for additional public health mission problems requiring automated decision support. The presentation will be based upon the underlying foundations of ontological representation as related to decision support, public health case detection and classification, and message validation. This session presents the use of enterprise knowledge services in support of the mission of public health. The objective of this workshop is to educate attendees with the technology, concepts, and uses of the public health ontology in automated reasoning systems. The goal of this workshop is to promote use and adoption of the public health reference ontology, ontology services, and reasoning frameworks based upon these services.. Each of the workshop sessions will provide an overview of the value propositions provided by this emerging ontology driven model paradigm to the community from a technical, public health domain and business perspective.

Summary

Summary

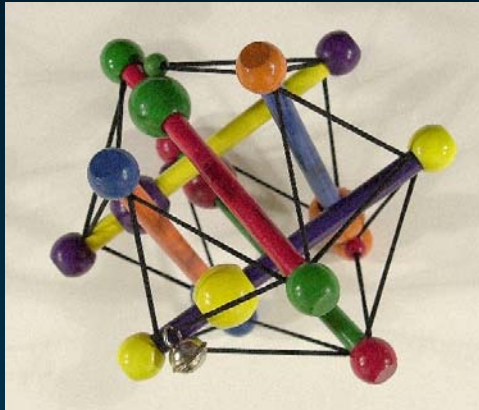
- ◆ Integration is key to making sense of the many health data repositories
 - E.g., Biosurveillance
- ◆ Two major data integration approaches
 - Data warehousing
 - Mediation
- ◆ Ontologies play a major role in data integration
 - Normalization of vocabulary
 - Supporting aggregation and inference

Current trends

- ◆ Semantic Web technologies provide a convenient platform for integrating biomedical data
 - Standard tooling
- ◆ Many biomedical ontologies available, ongoing harmonization
 - Standard vocabulary
- ◆ Emerging semantic interoperability specifications (e.g., HITSP, BRIDG, caBIG, HL7, ...)
 - Standard information models

Some persisting challenges

- ◆ Reconcile data annotated to different ontologies
 - Incomplete terminology integration systems (e.g., UMLS, RxNorm), limited harmonization between terminologies
 - Lack of permanent identifiers for biomedical entities
- ◆ Biomedical ontologies
 - Availability
 - Formalism (OWL, OBO, RRF, ...)
 - Quality



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