



Biobanks: Interoperability, Database and Ontology
November 23, 2010

Biobanks in the US
Interoperability, Database and Ontology



Olivier Bodenreider, MD, PhD

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

Outline

- ◆ Context in the US
 - Translational research
 - Biobanking
- ◆ Informatics infrastructure for translational research and biobanking
- ◆ Ontological aspects of biobanking
- ◆ Application – eMERGE Network



Translational research in the US

Translational medicine/research

◆ Definition

[Butte, JAMIA 2008]

- Effective transformation of information gained from biomedical research into knowledge that can improve the state of human health and disease

◆ Goals

- Turn basic discoveries into clinical applications more rapidly (“bench to bedside”)
- Provide clinical feedback to basic researchers



Flavors of translational research

◆ T1

- Move a basic discovery into a candidate health application
- “bench to bedside”

◆ T2

- Assess the value of T1 application for health practice leading to the development of evidence-based guidelines
- “clinical research”

◆ T3

- Move evidence-based guidelines into health practice, through delivery, dissemination, and diffusion research

◆ T4

- Evaluate the "real world" health outcomes of a T1 application in practice

http://medicalcenter.osu.edu/research/translational_research/Pages/index.aspx



Translational bioinformatics

- ◆ “... the development of storage, analytic, and interpretive methods to optimize the *transformation of increasingly voluminous biomedical data* into proactive, predictive, preventative, and participatory health.

*Translational bioinformatics includes research on the development of novel techniques for the **integration of biological and clinical data** and the evolution of clinical informatics methodology to encompass biological observations.*

*The end product of translational bioinformatics is **newly found knowledge** from these integrative efforts that can be disseminated to a variety of stakeholders, including biomedical scientists, clinicians, and patients.”*

AMIA strategic plan

<http://www.amia.org/inside/stratplan>



Enabling translational research

Clinical Translational Research Awards
(CTSA)

Translational research NIH Common Fund

Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI) National Institutes of Health • U.S. Department of Health and Human Services



The NIH Common Fund

Search: [GO](#)

<http://nihroadmap.nih.gov/>

Home Common Fund Programs Funding Opportunities Funded Research News & Events **About the Common Fund**

Back to: [Common Fund Home](#)

About the NIH Common Fund

The NIH Common Fund was enacted into law by Congress through the 2006 NIH Reform Act to support cross-cutting, trans-NIH programs that require participation by at least two NIH Institutes or Centers (ICs) or would otherwise benefit from strategic planning and coordination. The requirements for the Common Fund encourage collaboration across the ICs while providing the NIH with flexibility to determine priorities for Common Fund support.

To date, the Common Fund has been used to support a series of short term, exceptionally high impact, trans-NIH programs known collectively as the [NIH Roadmap for Medical Research](#). As the Common Fund grows, and research opportunities and needs emerge in the scientific community, the portfolio of programs supported by the Common Fund will likely evolve to encompass a diverse set of trans-NIH programs, although the NIH Roadmap is likely to remain a central component.

The Common Fund, including the programs of the NIH Roadmap for Medical Research, is coordinated by the [Office of Strategic Coordination](#), one of the five offices of the [Division of Program Coordination, Planning, and Strategic Initiatives \(DPCPSI\)](#) within the Office of the Director.

ARRA Information

- [View ARRA Funded Projects](#)
- [Search ARRA Funded Projects](#)



Budget, Planning and Legislative

- [Legislation](#)
- [Common Fund Strategic Planning Report 2009](#)



Clinical and Translational Science Awards

Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI) National Institutes of Health • U.S. Department of Health and Human Services

 The NIH Common Fund

Search: [GO](#)

<http://nihroadmap.nih.gov/>

Home **Common Fund Programs** Funding Opportunities Funded Research News & Events About the Common Fund

Back to: [Common Fund Home](#) > [Programs](#)

Clinical and Translational Science Awards (CTSAs)

- ▶ Overview
- ▶ [Implementation Group Members](#)
- ▶ [Funding Opportunities](#)
- ▶ [Funded Research](#)
- ▶ [Press Releases](#)

OVERVIEW

Growing barriers between clinical and basic research, along with the ever the increasing complexities involved in conducting clinical research, are making it more difficult to translate new knowledge to the clinic – and back again to the bench. These challenges are limiting professional interest in the field and hampering the clinical research enterprise at a time when it should be expanding.

Through discussions with deans of academic health centers, recommendations from the Institute of Medicine, and meetings with the research community, the NIH recognized that a broad re-engineering effort is needed to create greater opportunity to catalyze the development of a new discipline of clinical and translational science. The outcome, a bolder transforming vision for 21st Century, resulted in the launch of the **Clinical and Translational Science Awards (CTSA) Consortium** in October 2006.



Clinical and Translational Science Awards

◆ Strategic goals

- Build National Clinical and Translational Research Capability
- Provide Training and Improving the Career Development of Clinical and Translational Scientists
- Enhance Consortium-Wide Collaborations
- Improve the Health of our Communities and the Nation
- Advance T1 Translational Research

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



CTSA program

CTSA Clinical & Translational[®]
Science Awards

- ◆ National Center for Research Resources (NCRR)
- ◆ 55 academic health centers in 28 states (2010)
 - 60 centers upon completion
- ◆ Funding provided for 5 years
- ◆ Total annual cost: \$500 M
- ◆ Annual funding per center: \$4-23 M
 - Depending on previous funding

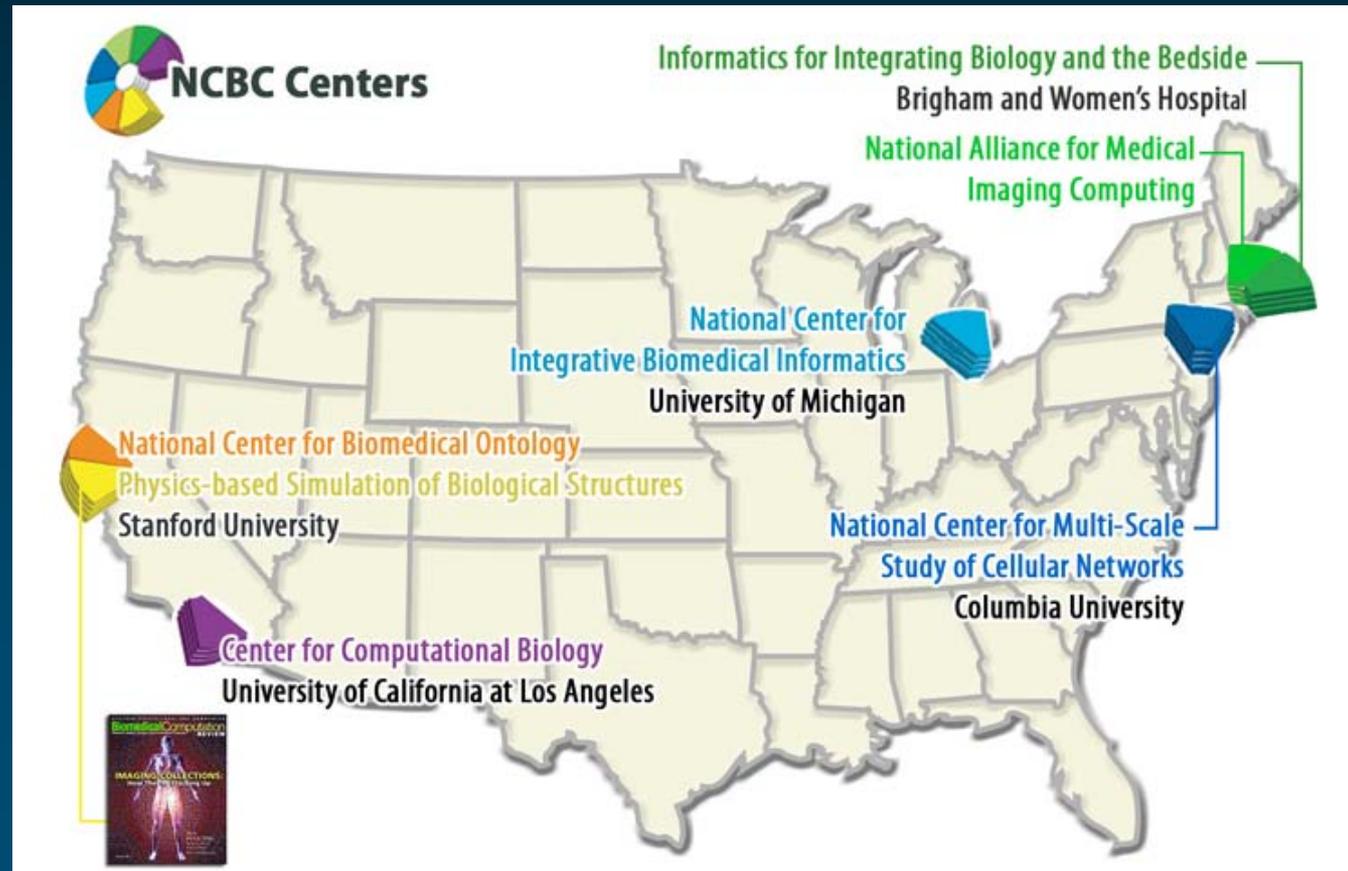
http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/



Other related programs

◆ National Centers for Biomedical Computing

“networked national effort to build the computational infrastructure for biomedical computing in the nation”



Other related programs



◆ Cancer Biomedical Informatics Grid (caBIG)

“an information network enabling all constituencies in the cancer community – researchers, physicians, and patients – to share data and knowledge.”

- Key elements
 - Bioinformatics and Biomedical Informatics
 - Community
 - Standards for Semantic Interoperability
 - Grid Computing
- 1000 participants from 200 organizations
- Funding: \$60 M in the first 3 years (pilot)

<https://cabig.nci.nih.gov/>



Biobanking in the US

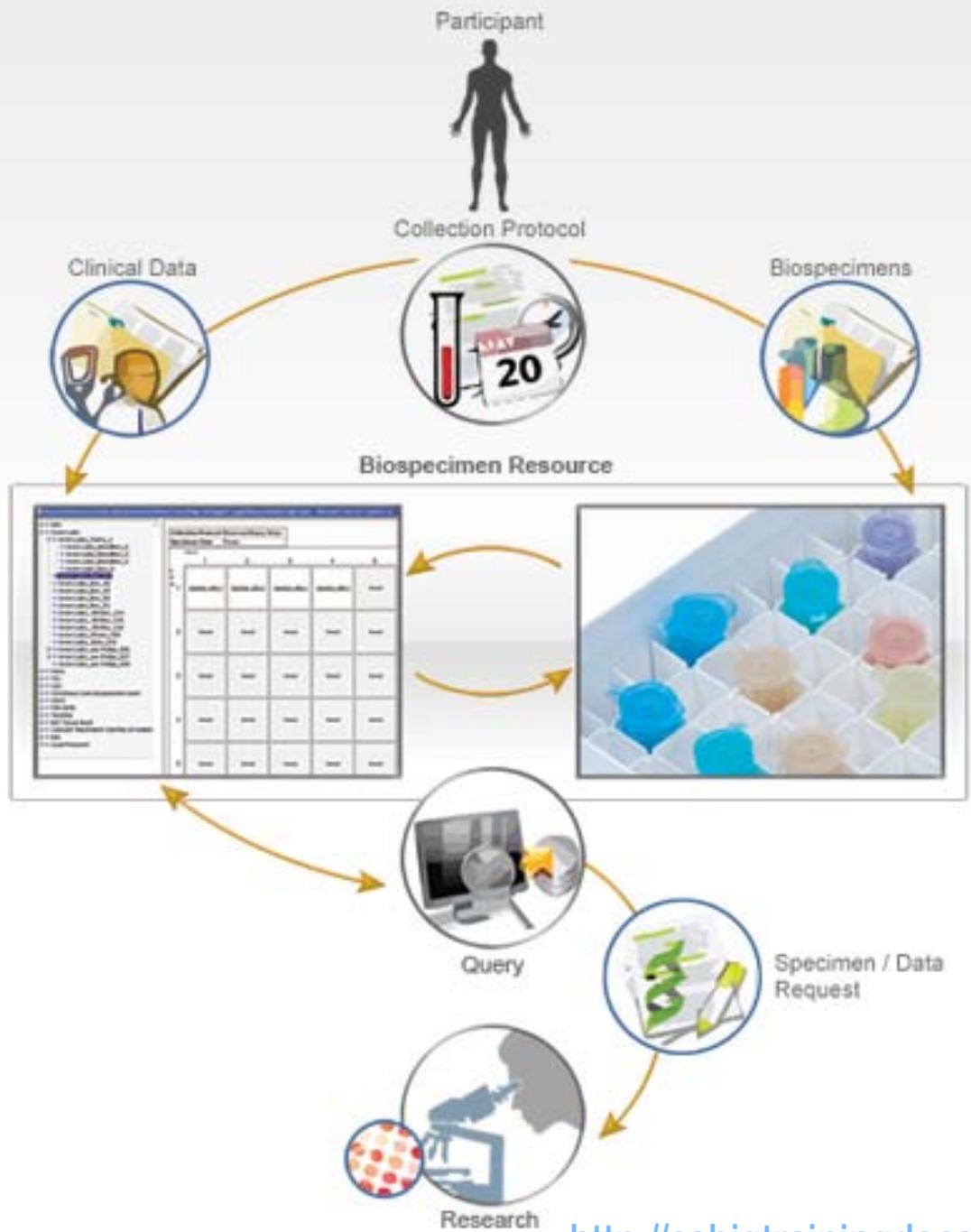
Biobank Definition

- ◆ Collection of **human tissue samples** and **medical information about donors**, which are stored for long periods of times and are used for research studies

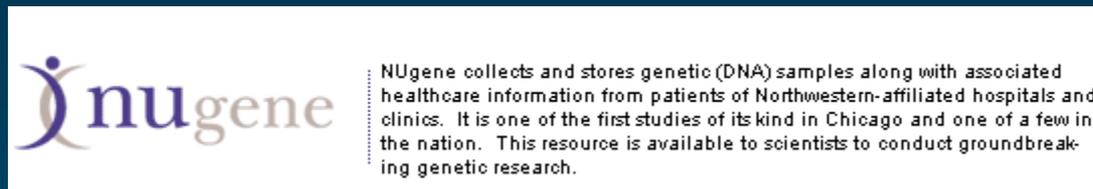
<http://www.biobank.org/>

- ◆ Bioregistry vs. Biorepository
 - Registry – collection of contact information
 - Biorepository – physical collection of samples





Examples of biobanks at large institutions



Genetic Alliance Biobank



Genetic Alliance BioBank

Infrastructures for
Solutions - **NOW!**

About BioBank

Members

Resources/Training

Join

Researchers

Home

Home

Welcome to
Genetic Alliance BioBank

**AN ADVOCACY OWNED REPOSITORY
FOR BIOLOGICAL SAMPLES AND CLINICAL DATA**



Missed the webinar,
Biobank Governance?

[Click here to view the
archive.](#)

Genetic Alliance Registry and BioBank is a centralized, clinical data registry and sample repository that enables translational genomic research. Founded in 2003 by leaders in disease research advocacy field, this cooperative venture provides shared infrastructure and customized solutions for disease advocacy organizations to lead sophisticated research initiatives.

What's New

- [Disease-Specific Organizations Invited to Apply for \\$20,000 BioBank Grant](#)
- [Executive Summary](#)
- [Private Access and Genetic Alliance Announce Strategic](#)

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



CTSA Biobank Consortium

CTSA Biobank Consortium

The University of Texas Health Science Center at Houston is leading the CTSA Biobank Consortium, a collaboration to harmonize policies and procedures and to embed regulatory compliance in support of a nationwide ability to locate, request, and receive samples and sample-related clinical data. The Biobank will provide researchers with 1) the ability to search for and locate samples and 2) a standardized, simplified process for requesting samples on the basis of demographic and patient consent information, thereby eliminating many of the existing hurdles that prevent substantive genomic research from taking place.

The Biobank employs a federated model that encourages participation by sample owners who are concerned about guaranteeing ownership of sample and clinical data. In a federated model, individual sites agree on shared policies and procedures for data and sample sharing as well as oversight. Samples remain with and are governed by the contributing investigator at each site. The contributing investigator has final authority on whether to collaborate with or release samples to qualified researchers.

The Biobank will create an informatics system that works with the established software and laboratory procedures at all CTSA sites. Currently, the internal prototype allows queries to be made to participating sites. The project team anticipates an application release to allow researchers to make actual requests for samples by 2013.

<http://www.ncrr.nih.gov/ctsa/newsletter/December2009/>



[About caHUB](#) | [caHUB Development](#) | [Partner With Us](#) | [News Center](#) | [Resources](#)

Overview

The National Cancer Institute (NCI) is sponsoring a unique initiative to advance cancer research and treatment through development of the cancer Human Biobank (caHUB), a national biorepository of human tissue, blood, other biological materials, and a comprehensive online database.

caHUB was created in response to the critical and growing shortage of high-quality, well-documented biospecimens for development of molecularly based diagnostic and therapeutic agents that will further enable personalized treatment for cancer patients.

The caHUB initiative builds on resources already developed by the NCI, including the [Biospecimen Research Network](#) and the [NCI Best Practices for Biospecimen Resources](#), both of which were developed to address challenges around standardization of the collection and dissemination of quality biospecimens.



Office of Biorepositories and Biospecimen Research (OBBR)

<http://cahub.cancer.gov/>



caHUB Biospecimen Research Network

◆ Missions

- Bridging the gap between existing clinical practice for biospecimens and emerging technologies for personalized diagnostics and therapies
- Defining the **most significant variables** for prospective collection of tissues, blood, and bodily fluids
- Developing evidence-based **biospecimen quality indicators** for specific analytical platforms

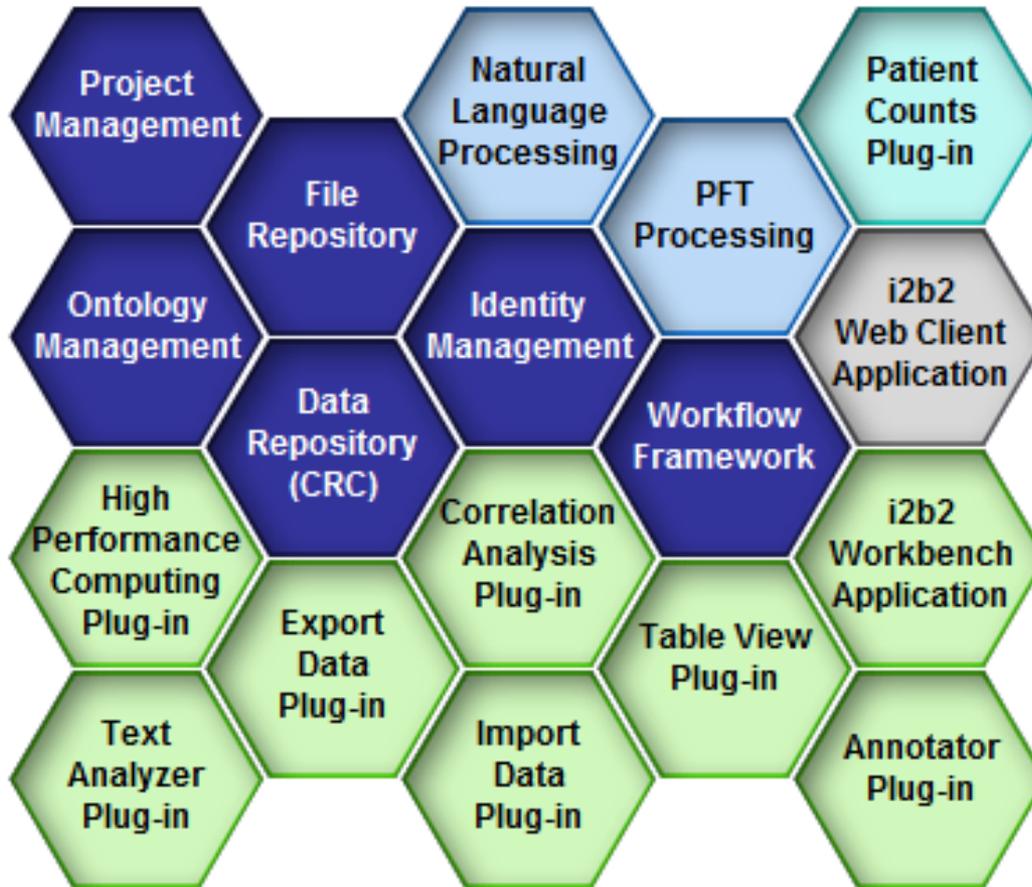
<http://biospecimens.cancer.gov/researchnetwork/>



Informatics infrastructure

i2b2 software

i2b2 software



i2b2 Hive

i2b2 software

- ◆ Developed by the i2b2 NCBC (Boston)
- ◆ Adopted/adapted by most CTSA sites
- ◆ Clinical research data warehouse
- ◆ No specific features for biospecimens
- ◆ Open source, customizable
- ◆ Selected by the CTSA Biobank Consortium

<https://www.i2b2.org/software/>



caTissue

- ◆ Biorepository management tool designed for biospecimen inventory, tracking, and annotation
- ◆ Developed by Washington University in St. Louis
- ◆ Part of caBIG
- ◆ Integration through caGrid
- ◆ Coupled with caTIES – cancer Text Information Extraction System
- ◆ Adopted by several academic centers



Ontological aspects of biobanking

Genetic Alliance Biobank TRIMS

- ◆ TRIMS is a centralized data repository for annotation data (clinical and experimental), sample source and handling information, processing and quality assurance (QA) information, as well as inventory and process flow data.
- ◆ All clinical, demographic, and experimental data management and sample registration/accessioning occur in TRIMS. TRIMS also provides tracking, data query, report generation, and process management functions.
- ◆ [...]
- ◆ TRIMS uses **SNOMED**, from the College of American Pathologists, as the controlled vocabulary for classification of disease and sample morphology.
- ◆ TRIMS also uses **internal standards** for controlling the quality and consistency of entered data including use of drop-down pick lists, flexible units translation, and field data type restrictions among others.
- ◆ [...]

<http://www.biobank.org/english/View.asp?x=1413>



Ontologies

- ◆ Provide the vocabulary for the annotation of clinical datasets and biospecimens
- ◆ Clinical annotations
 - Diseases
 - Pathological conditions
- ◆ Biospecimens
 - Localization (anatomical, subcellular)
 - Collection procedure

Ontologies Clinical annotations

- ◆ EMRs
 - SNOMED CT
- ◆ Billing data
 - ICD-9-CM
- ◆ Clinical research repositories
 - NCI Thesaurus
- ◆ Registries
 - ICD-O



Ontologies Biospecimen annotations

<http://www.brenda-enzymes.info/>

◆ Anatomy

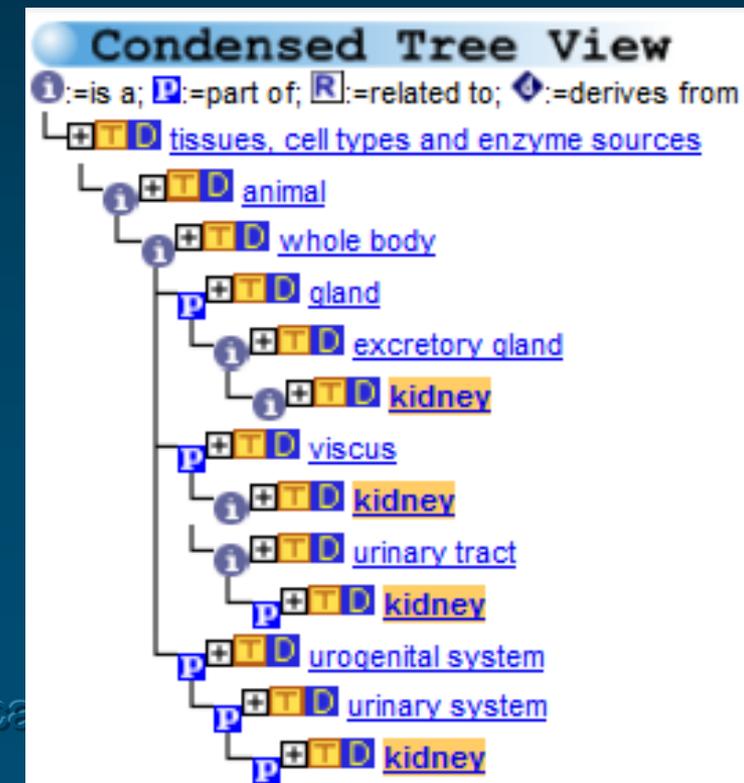
- Foundational Model of Anatomy (FMA)
- Anatomical component of general ontologies (SNOMED CT, NCI Thesaurus)

◆ Tissue/cell ontologies

- BRENDA (OBO)
- Cell type ontology (OBO)

◆ Morphology (cancer)

- ICD-O



Ontology for Biomedical Investigations

- ◆ Domain poorly covered by most ontologies
- ◆ International collaborative effort (OBI Consortium) – OBO Family



<http://obi-ontology.org/>

- ◆ In the spirit of the MGED ontology for microarray experiments
- ◆ Scope
 - biological material – blood plasma
 - instrument (and parts of an instrument therein) – DNA microarray, centrifuge
 - information content (e.g., image) or digital information entity (e.g., electronic medical record)
 - design and execution of an investigation (and individual experiments therein) – study design, electrophoresis material separation
 - data transformation – principal components analysis, dimensionality reduction, mean calculation

Ontology repositories

◆ Unified Medical Language System

- National Library of Medicine
- Free (license agreement required)
- ~ 100 ontologies
- 2.2 M concepts
- Annotation tool (MetaMap)

<http://umlsks.nlm.nih.gov/>



◆ BioPortal

- National Center for Biomedical Ontology
- Free
- ~ 202 ontologies
- 1.4 M concepts
- Annotation tool (Annotator)

<http://bioportal.bioontology.org/>

Terminology integration systems

- ◆ UMLS and BioPortal
- ◆ Enable mappings across terminologies
- ◆ Useful for reconciling annotations to disparate terminologies

Some issues with ontologies

- ◆ Information model required
 - Ontologies provide vocabulary for data elements and categorical values
- ◆ Limited coverage
 - Biobanking procedures
- ◆ Availability
 - SNOMED CT – available only to the members of the IHTSDO
- ◆ Limited availability of coded EMR data
 - Need for natural language processing of EMR data



Application

eMERGE Network



eMERGE Network Overview

- ◆ Funded by NIH
- ◆ 5 member sites
 - Group Health Cooperative / University of Washington
 - Marshfield Clinic
 - Mayo Clinic
 - Northwestern University
 - Vanderbilt University
- ◆ Local biorepositories
- ◆ Explore whether EMR systems can serve as resources for complex genomic analysis of disease susceptibility and therapeutic outcomes, across diverse patient populations

https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page



eMERGE Network Methods

◆ Primary phenotypes studied

- Dementia - GHC/University of Washington
- Cataract - Marshfield Clinic
- Peripheral Arterial Disease - Mayo Clinic
- Type 2 Diabetes - Northwestern University
- Cardiac conduction - Vanderbilt University

◆ Tooling developed

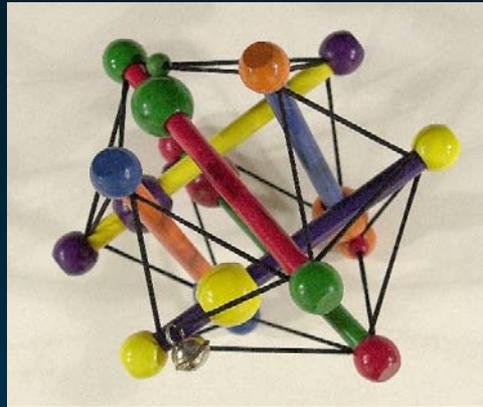
- eleMAP – facilitate mapping of local phenotype data dictionaries to standard terminologies
- PheWAS



eMERGE Network Results

- ◆ Demonstrated the feasibility of using EMR data for identifying genotype-phenotype associations (e.g., “rediscovery” of association between genes and red blood cell traits)
- ◆ Demonstrated the feasibility of pooling clinical data from various institutions, increasing statistical power
- ◆ Confirmed the need for NLP analysis of EMR data (clinical notes, reports)





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