caBIG® in the Trenches: Deploying an Infrastructure that Enables Collaboration

R. Mark Adams, Ph.D.
Principal
Booz Allen Hamilton

April 21, 2010

Key Concepts

- 21st century scientific research requires new models of collaboration and technology that enables data interoperability
- Widely-recognized data standards, and technologies that leverage them are critical for data interoperability
- These technologies enable a fundamentally new type of scientific communication
- caBIG® tools, standards and technology supports a wide range of biomedical research activities
- Use of caBIG® is widely supported through a diverse collection of government, academic, and commercial sources

Agenda

- The Need for Collaboration
- An Overview of caBIG®
- Services and Interoperability
- caBIG® Enables Collaboration
- Getting Started with caBIG®

The Need for Collaboration

Collaboration as a Means to Discovery

- Drivers for collaborative research
  - Pre-competitive space for drug discovery and development continues to grow
  - Volume of high-quality, publicly-available data continues to increase
  - Research associated expenses continue to increase
  - New models of drug discovery continue to evolve

- Novel discoveries increasingly rely on multi-disciplinary team research

Information Exchange: Yesterday AND Today

17th century:

Royal Society of London

- Oldest learned society (1660)
- Oldest scientific journal (1665)

21st century:

- •Nature, Science, Cancer Research, Journal of Clinical Oncology magazines and journals

Science is Increasingly Driven by Information Sharing and Collaboration

Examples of collaborative Science

- GenBank – driving the genomics revolution
- PDB – enabling rational drug design
- Array Express – fueling functional genomics

Barriers to Collaborative Research

- Tsunami of Genomic and Clinical Data
- Islands of Information
- Standard Language
- IT Systems Do Not Interoperate

An Overview of caBIG®

caBIG®: Biomedical Information Highway

The cancer Biomedical Informatics Grid® (caBIG®) is a virtual network of interconnected data, individuals, and organizations that redefines how research is conducted, care is provided, and patients/participants interact with the biomedical research enterprise.

caBIG® is a Path to Overcome Obstacles

- Tsunami of Genomic and Clinical Data
  - 40+ Software Tools
- Islands of Information
  - National Grid for Data Sharing
- Standard Language
  - Standardized Vocabularies
- IT Systems Do Not Interoperate
  - Interoperability

caBIG® Capabilities Enable

Discovery > Translation > Clinical Research

Molecular Medicine

- Clinical Research
  - Track clinical trial registrations
  - Facilitate automatic capture of clinical laboratory data

    - Manage reports describing adverse events during clinical trials

- Molecular Biology
    - Combine proteomics, gene expression, and other basic research data
    - Submit and annotate microarray data
    - Integrate microarray data from multiple manufacturers and permit analysis and visualization of data
- Imaging
    - Utilize the National Cancer Imaging Archive repository for medical images including CAT scans and MRIs
    - Visualize images using DICOM-compliant tools
    - Annotated Images with distributed tools
- Pathology
    - Access a library of well characterized, clinically annotated biospecimens
    - Use tools to keep an inventory of a user's own samples
    - Track the storage, distribution, and quality assurance of specimens

caBIG® Core Principles

- Open Access – caBIG® is open to all, enabling wide-spread access to tools, data, and infrastructure

- Open Development – Planning, testing, validation, and deployment of caBIG® tools and infrastructure are open to the entire research community

- Open Source – The underlying software code of caBIG® tools is available for use and modification

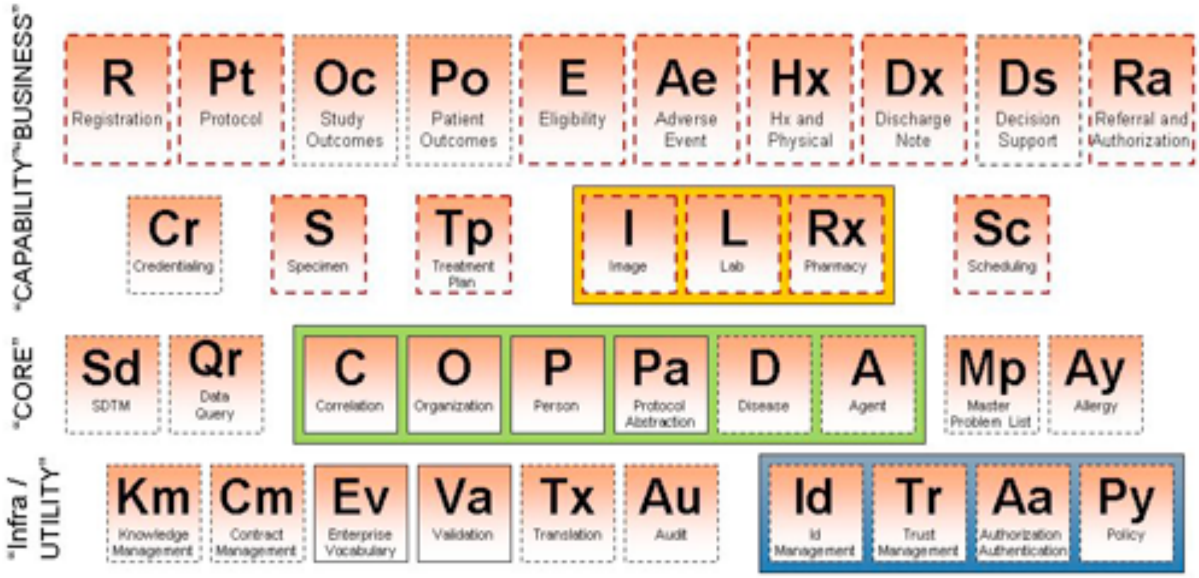
- Federation – Resources can be controlled locally, or integrated across multiple sites

Services and Interoperability

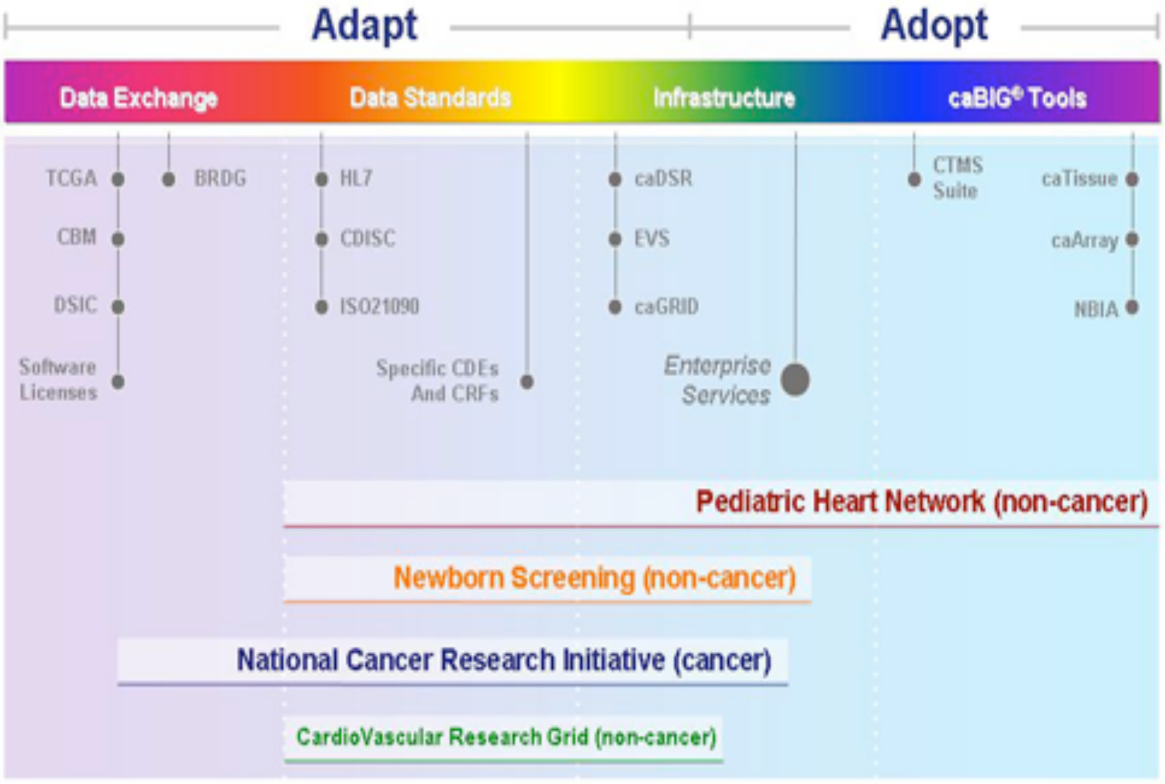
Boundaries and Interfaces

- Focus on boundaries and interfaces, how things fit together, NOT on the internal details
- Once they're built: assume inner details will be diverse & changing

caBIG® Periodic Table of Services



Interoperability Spectrum



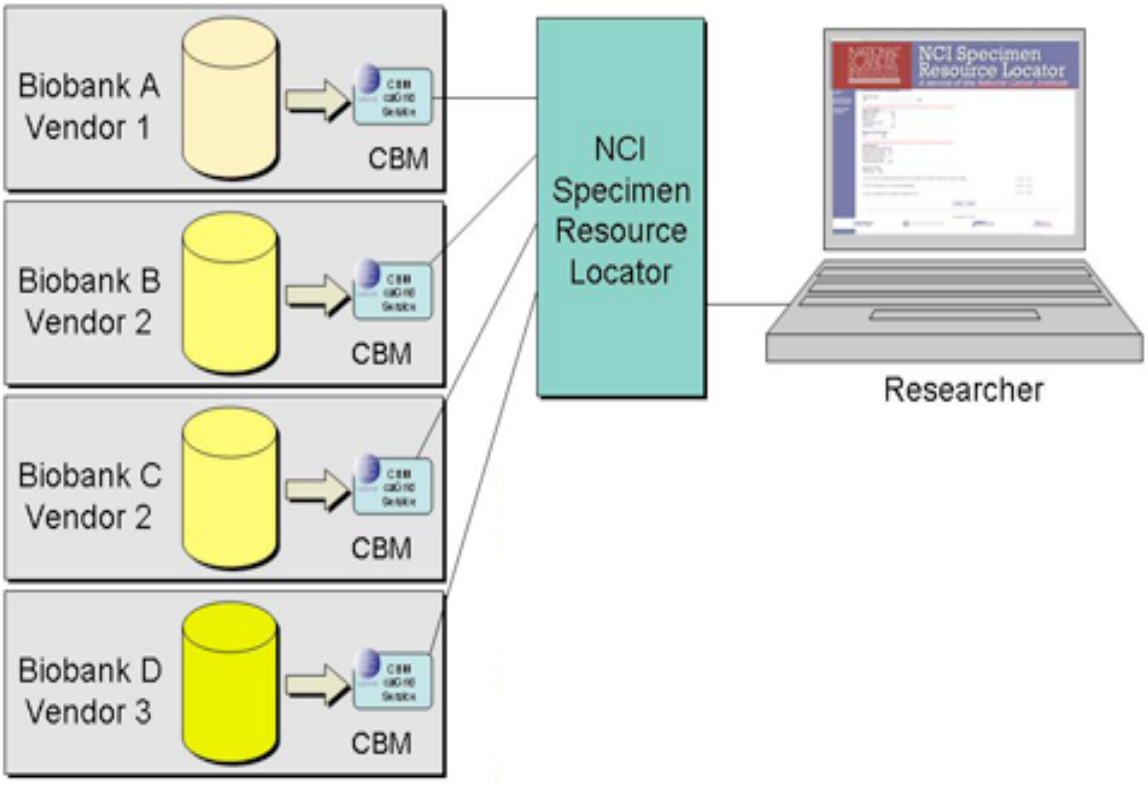
Open Source in Standards

- Open access to core tools promotes rapid adoption of standards
- Standards are perceived as having “shared ownership” within the community, instead of being property of a single company

- Standards can be tested by widespread use in many different contexts, allowing for simplified validation activity.
- Everyone (except for some vendors...) benefits from widespread adoption of tools for data access.

Services Example: Common Biorepository Model (CBM)

Provides an easy path for biobanks to share their data:



- Summary level data about biospecimen collections
- Key use-case is finding specimens- NCI Specimen Resource Locator

Software Vendors Participating in CBM Development and Testing (April 2010)

- Artificial Intelligence in Medicine (AIM)
- Aptia Systems
- BioFortis
- caTissue suite
- Daedualus Software
- Freezerworks
- Genvault

- GenoLogics (GLS)
- HealthCare IT, Inc.
- IMS, Inc.
- LabVantage
- Ocimum Biosolutions
- PercipEnz
- ThermoFisher Scientific
- Waban Software

caBIG® Enables Collaboration

caBIG®: a Growing Community…

- More than 2300 individuals from 740+ institutions
- 56 NCI-designated Cancer Centers
- 18 NCI Community Cancer Centers
- 1100+ Attendees at 2009 caBIG® Annual Meeting
- 10 Workspaces (18 Special Interest Groups)
- 6 Knowledge Centers (13 Organizations)
- Commercial Service Providers (15 licensed companies)

Organizations Participating Include*…

- Abbott Laboratories
- Astra Zeneca
- Cardiff University (UK)
- Center for the Development of Advanced Computing (CDAC – IN)
- Centocor
- Curie Institute (FR)
- Dublin Institute of Technology (IR)
- Drexel University
- Eli Lily
- Erasmus Medical Center (NL)
- FDA
- Friedrich Miescher Institute for Biomedical Research (CH)
- Genentech
- Genesis R&D Inc (NZ)
- Glaxo Smith Kline
- Hiroshima University (JP)
- Imperial College of London (UK)
- INSERM (FR)
- Medarex

- Moscow State University (RU)
- National University of Singapore (SG)
- National Yang-Ming University (TW)
- Ontario Cancer Institute (CA)
- Pune University (IN)
- Queensland University (AU)
- Roche Holding AG (DE)
- Taiho Pharmaceutical Co., Ltd. (JP)
- Takeda
- Tulane University
- University of Crete (CR)
- University of Edinburgh (UK)

caBIG® is Establishing Global Connections

- United States
- Mexico
- Chile
- Uruguay
- Argentina
- Brazil
- UK
- Netherlands
- Germany
- Czech Republic
- Finland
- Jordan
- India
- China
- Australia
- New Zealand

caBIG®, a biomedical research "highway", connecting a growing number of people and organizations across the globe

caBIG® Examples of Success

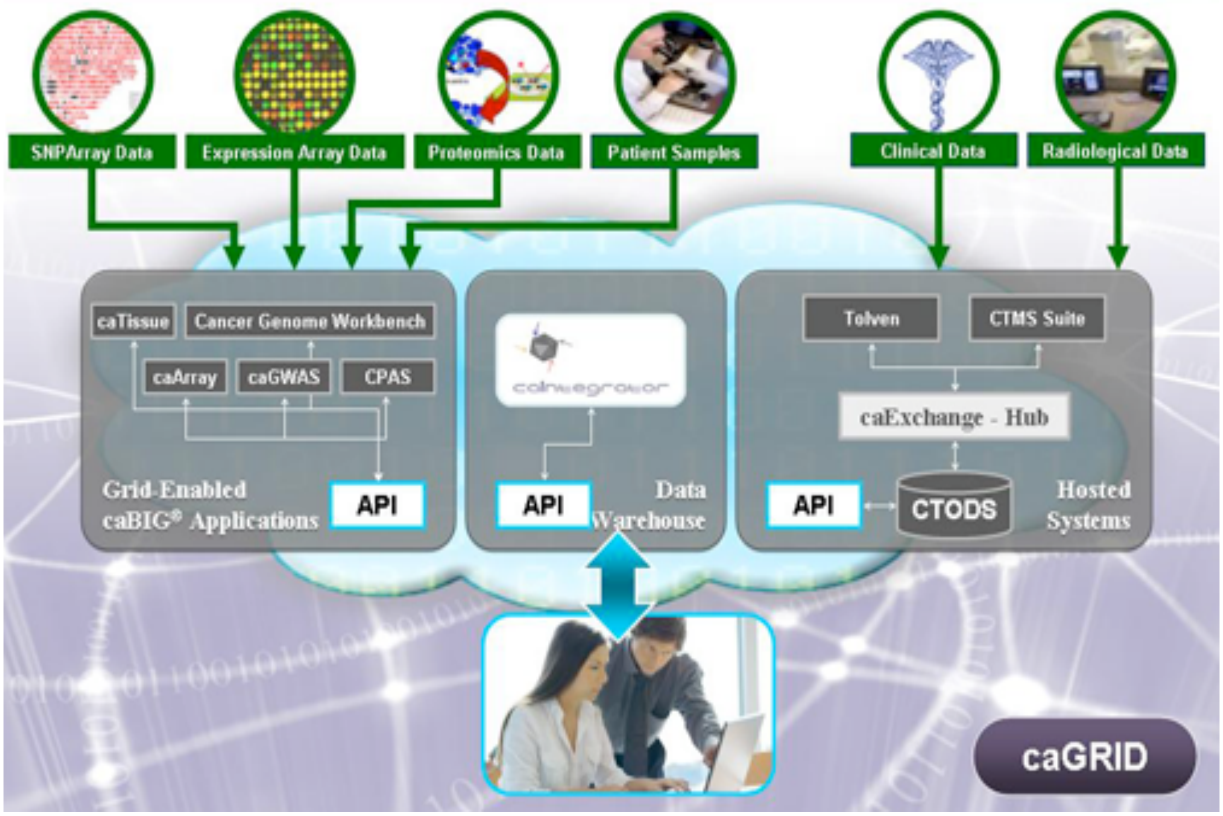
- Washington University at St Louis
  - Hosting 450K biospecimens on caGrid
  - Developed interoperable clinical data warehouse spanning 13 hospitals in the system
- The Ohio State University
  - Using caGrid to connect University and city hospital in Ohio Perinatal Research Project
  - Created federated, searchable repositories for clinical trial metadata with CTSA-funded sites
- University of Alabama, Birmingham

- Applying caGrid technology to connect diverse collection of legacy IT systems, including billing, radiology and clinical records across the university

I-SPY Trial: Identify Biomarkers Predictive of Therapeutic Response in Stage 3 Breast Cancer

- Multiple Morphologic Patterns of Breast Cancer
- Multiple Sites/Organizations
  - Specialized Programs of Excellence (SPOREs)
  - Cancer and Leukemia Group B (CALGB)
  - American College of Radiology Imaging Network (ACRIN)
  - University of California at San Francisco (UCSF)
- Multiple Data Types
  - Clinical diagnosis
  - Treatment history
  - Histologic diagnosis
  - Pathologic status
  - Tissue anatomic site
  - Surgical history
  - Gene expression
  - Chromosomal copy number
  - Loss of heterozygosity
  - Methylation patterns
  - miRNA expression
  - DNA sequence

I-SPY Trial IT Infrastructure



Using caBIG® to Classify Lymphoma

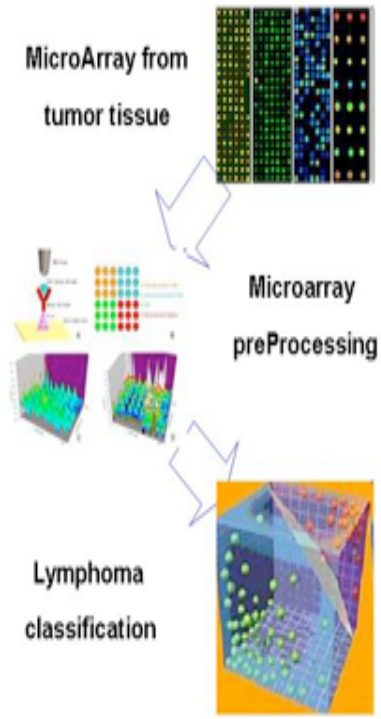
•Scientific value
- Use gene-expression patterns associated with two lymphoma types to predict the type of an unknown sample.
- Connect caGrid data service (caArray) with analytical services (PreProcess, SVM and KNN from GenePattern).

•Major steps
- Querying training data from experiments stored in caArray.
- Preprocessing, i.e., normalizing the microarray data.
- Predicting lymphoma type using SVM & KNN services.

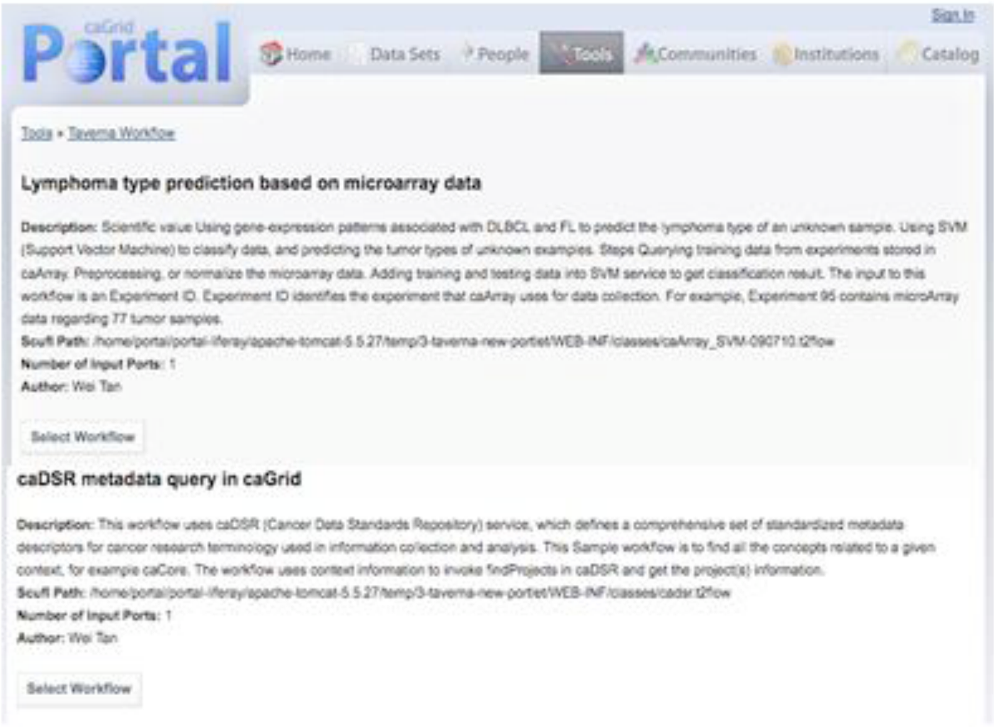
•Extension
- Generalized the workflow into a cancer type prediction routine that can be used on other caArray data sets.

Ravi Madduri, Univ. Chicago
Carole Goble, U. Manchester, UK
Wei Tan, Univ. Chicago
Dinanath Sulakhe, Univ. Chicago
Stian Soiland-Reyes, U Manchester, UK

Lymphoma Prediction Workflow



ck. Juli Klemm, Xiaopeng Bian, Rashmi Srinivasa (**NCI**), Jared Nedzel (**MIT**)



Tools » Taverna Workflow

Lymphoma type prediction based on microarray data

Description: Scientific value Using gene-expression patterns associated with DLBCL and FL to predict the lymphoma type of an unknown sample. Using SVM (Support Vector Machine) to classify data, and predicting the tumor types of unknown examples. Steps Querying training data from experiments stored in caArray. Preprocessing, or normalize the microarray data. Adding training and testing data into SVM service to get classification result. The input to this workflow is an Experiment ID. Experiment ID identifies the experiment that caArray uses for data collection. For example, Experiment 95 contains microArray data regarding 77 tumor samples.
Scufl Path: /home/portal/portal-liferay/apache-tomcat-5.5.27/temp/3-taverna-new-portlet/WEB-INF/classes/caArray_SVM-090710.t2flow
Number of Input Ports: 1
Author: Wei Tan

Select Workflow

caDSR metadata query in caGrid

Description: This workflow uses caDSR (Cancer Data Standards Repository) service, which defines a comprehensive set of standardized metadata descriptors for cancer research terminology used in information collection and analysis. This Sample workflow is to find all the concepts related to a given context, for example caCore. The workflow uses context information to invoke findProjects in caDSR and get the project(s) information.
Scufl Path: /home/portal/portal-liferay/apache-tomcat-5.5.27/temp/3-taverna-new-portlet/WEB-INF/classes/cadsr.t2flow
Number of Input Ports: 1
Author: Wei Tan

Select Workflow

caBIG® Enables Translational Research Beyond Cancer

- Pediatric Heart Network (PHN)
  - caBIG® infrastructure (caGrid, LexEVS) and applications (NBIA, caTissue) can connect pediatric researchers and enable secure data sharing
- National Institute of Child Health and Human Development (NICHD) Pediatric Terminology Initiative
  - caBIG® tools (NCI Thesaurus, caDSR) help manage metadata produced by the program
- NCI Mouse Models of Human Cancer
  - Using NBIA and other caBIG® tools in support of mouse genetics research

NICHD Pediatric Terminology Project

- In January 2009, the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) charged a small team with beginning the process of developing a framework and process for harmonizing terminology for pediatric research
- The process for harmonization of content specific data acquisition tools includes:
  - Harmonizing and vetting of tool concepts by the stakeholder community
  - Development of harmonized terminology to be used by the research community

Project Objective: To develop a harmonized terminology system for pediatrics and pediatric conditions thereby establishing a basis for enabling semantically unambiguous data sharing, allowing aggregation and comparison of data collected at different times or by different groups, resulting in richer analyses

Terminology Development Process: Scientific Resources, Modeling Technology, Semantic Infrastructure, and Open Source Tools

- Sources
  1. Trace list of sources
  2. Draft tool
  3. Structure concepts
  4. Develop model
  5. Curate Common Data Elements
  6. Generate Research Tool

- Final Examination Tool
- Common Data Element Browser
- UML Model
- NCI Thesaurus
- Draft Examination Tool

Getting Started with caBIG®

caBIG® - Diverse Support Channels

- caBIG Program Support Regularly-scheduled in-person workshops, webinars, and training sessions at national meetings (ASCO, AACR, BioIT), as well as on-lineTutorials and Videos and Learning Center materials

- Knowledge Centers (https://cabig.nci.nih.gov/esn/knowledge_centers) serve as the nexus for an expanding community employing caBIG® tools, standards, and infrastructure in a specific domain. Knowledge Center staff can provide expert guidance to end users, IT staff and senior decision makers implementing caBIG® tools and infrastructure.

- Support Service Providers (https://cabig.nci.nih.gov/esn/service_providers) are able to provide comprehensive technical support under client-specific agreements. There are four categories of services offered by caBIG® Support Service Providers:
  - Help Desk Support
  - Adaptation and Enhancement of caBIG®-Compatible Software
  - Deployment Support for caBIG® Software Applications
  - Documentation and Training Materials and Services

Finding What You Need…

- If you are a basic researcher
  - https://cabig-kc.nci.nih.gov/Molecular/KC/
- If you are a clinical researcher
  - https://cabig-kc.nci.nih.gov/CTMS/KC/
- If you are interested in biospecimen management
  - https://cabig-kc.nci.nih.gov/Biospecimen/KC/
- If you are a software developer and want technical information
  - https://cabig.nci.nih.gov/
- If you have questions about a specific software application
  - http://ncicb.nci.nih.gov/support

Finding What You Need…

- If you want additional general information about caBIG®
  - http://cabig.cancer.gov/
- If you want to receive our monthly e-newsletter
  - http://cabig.cancer.gov/resources/newsletter/
- If you want a complete overview of the caBIG® program
  - https://cabig.nci.nih.gov/training/cabigessentials/player.html
- If you want a complete list of caBIG® tools
  - https://cabig.nci.nih.gov/adopt/
- If you want a demo-for-the-perplexed
  - Call (301) 594-3602

Take-Home Messages

- *21st century scientific research requires new models of collaboration and technology that enables data interoperability*
- *Widely-recognized data standards, and technologies that leverage them are critical for data interoperability*
- *These technologies enable a fundamentally new type of scientific communication*
- *caBIG® tools, standards and technology supports a wide range of biomedical research activities*
- *Use of caBIG® is widely supported through a diverse collection of government, academic, and commercial sources*

Questions?