Vision and Infrastructure Behind the Cancer Biomedical Informatics Grid

Peter A. Covitz, Ph.D.
Director, Core Infrastructure
National Cancer Institute
Center for Bioinformatics
The Center for Bioinformatics is the NCI’s strategic and tactical arm for research information management.

- We collaborate with both intramural and extramural groups.
- Mission to integrate and harmonize disparate research data.
- Production, service-oriented organization. Evaluated based upon customer and partner satisfaction.
NCICB Operations teams

- Systems and Hardware Support
- Database Administration
- Software Development
- Quality Assurance
- Technical Writing
- Application Support and Training
- caBIG Management
National Cancer Institute 2015 Goal

Relieve suffering and death due to cancer by the year 2015
Origins of caBIG

- **Need:** Enable investigators and research teams nationwide to combine and leverage their findings and expertise in order to meet NCI 2015 Goal.

- **Strategy:** Create scalable, actively managed organization that will connect members of the NCI-supported cancer enterprise by building a biomedical informatics network.
Scenario from caBIG Strategic Plan

A researcher involved in a phase II clinical trial of a new targeted therapeutic for brain tumors observes that cancers derived from one specific tissue progenitor appear to be strongly affected.

The trial has been generating proteomic and microarray data. The researcher would like to identify potential biochemical and signaling pathways that might be different between this cell type and other potential progenitors in cancer, deduce whether anything similar has been observed in other clinical trials involving agents known to affect these specific pathways, and identify any studies in model organisms involving tissues with similar pathway activity.
caBIG Governance and Organization
caBIG Governance Models

Feudalism

Warlord culture offers little incentive to cooperate
Governance Models

Forced Collectivization

Centralized monolithic approach not flexible or scalable
Governance Models

Federal Democracy

Balance between central management and local control. Best fit for caBIG Principles.
caBIG Organization Structure

caBIG Oversight

General Contractor

Clinical Trial Mgmt
Integrative Cancer Research
Tissue Banks & Pathology Tools

Architecture

Vocabularies & Common Data Elements

Working Group

Working Group

Working Group

Working Group

Strategic Working Groups

= Project
Interoperability

ability of a system to access and use the parts or equipment of another system

Syntactic interoperability

Semantic interoperability
## caBIG Compatibility Guidelines

<table>
<thead>
<tr>
<th>Maturity Model</th>
<th>Legacy</th>
<th>Bronze</th>
<th>Silver</th>
<th>Gold</th>
</tr>
</thead>
</table>
| **Programming and Messaging Interfaces** | - No programmatic interfaces to the system are available. Only local data files in a custom format can be read  
- Data transfer mechanisms implemented only on an ad hoc basis | - Programmatic access to data from an external resource is possible. | - Well-described API’s provide access to data in the form of data objects.  
- Standards-based electronic data formats are supported for both input to and output from the system.  
- Standards-based messaging protocols are supported wherever messaging is relevant. | - All features of Silver, plus:  
- Service-oriented components produce or use resources in the form of grid services  
- Interoperable with data grid architecture to be defined by caBIG |
| **Vocabularies / Terminologies & Ontologies** | - Free text used throughout for data collection  
- All terminologies must include unambiguous definitions of terms | - Use of publicly accessible controlled vocabularies as well as local terminologies. | - Terminologies reviewed and validated by the caBIG Vocabulary/Common Data Element (VCDE) Workspace used for all appropriate data collection fields. | - All features of Silver, plus:  
- Full adoption of caBIG terminology standards as approved by the VCDE workspace |
| **Data Elements** | - No Structured metadata is recorded | - Data element descriptions are maintained with sufficient definitional depth to enable a subject matter expert to unambiguously interpret the contents of the resource without contacting the original investigator. | - Common Data Elements (CDEs) built from controlled terminologies and according to practices validated by the VCDE workspace are used throughout.  
- CDEs are registered as ISO/IEC 11179 metadata components in the cancer Data Standards Repository (caDSR) | - All features of Silver, plus:  
- CDEs designated as caBIG Standards by the VCDE workspace are used  
- Metadata is advertised and discoverable via the caBIG grid services registry |
| **Information Models** | - No model describing the system is available in electronic format | - Diagrammatic representation of the information model is available in electronic format. | - Information models are defined in UML as class diagrams and are reviewed and validated by the VCDE Workspace. | - All features of Silver, plus:  
- Information models are harmonized across the caBIG Domain Workspaces |
Model-Driven Architecture
OMG Model Driven Architecture

**How Systems Will Be Built**

MDA® provides an open, vendor-neutral approach to the challenge of business and technology change. Based firmly upon OMG’s established standards*, MDA aims to separate business or application logic from underlying platform technology. Platform-independent applications built using MDA and associated standards can be realized on a range of open and proprietary platforms, including CORBA®, J2EE, .NET, and Web Services or other Web-based platforms. Fully-specified platform-independent models (including behavior) can enable intellectual property to move away from technology-specific code, helping to insulate business applications from technology evolution, and further enable interoperability. In addition, business applications, freed from technology specifics, will be more able to evolve at the different pace of business evolution.

---

*Key standards that make up the MDA suite of standards include Unified Modeling Language (UML); Meta-Object Facility (MOF); XML Meta-Data Interchange (XMI); and Common Warehouse Meta-model (CWM).
MDA Approach

- Analyze the problem space and develop the artifacts for each scenario
  - Use Cases
- Use Unified Modeling Language (UML) to standardize model representations and artifacts. Design the system by developing artifacts based on the use cases
  - Class Diagram – Information Model
  - Sequence Diagram – Temporal Behavior
- Use meta-model tools to generate the code
Limitations of MDA

- Limited expressivity for semantics
- No facility for runtime semantic metadata management
caCORE

MDA plus a whole lot more!
caCORE

Bioinformatics Objects

Common Data Elements

Enterprise Vocabulary

SECURITY
Use Cases

- Description
- Actors
- Basic Course
- Alternative Course
Bioinformatics Objects
What do all those data classes and attributes actually mean, anyway?

Data descriptors or “semantic metadata” required

Computable, commonly structured, reusable units of metadata are “Common Data Elements” or CDEs.

NCI uses the ISO/IEC 11179 standard for metadata structure and registration

Semantics all drawn from Enterprise Vocabulary Service resources
<table>
<thead>
<tr>
<th>Concept Code</th>
<th>Relationships</th>
<th>Preferred Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2919</td>
<td>Disease_Has_Abnormal_Cell: Adenocarcinoma Cell</td>
<td>Prostate Adenocarcinoma</td>
<td>NCI</td>
</tr>
<tr>
<td></td>
<td>Disease_Has_Associated_Anatomic_Site: Male Reproductive System</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disease_Has_Associated_Anatomic_Site: Prostate Gland</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disease_Has_Normal_Cell_Origin: Glandular Cell</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disease_Has_Normal_Tissue_Origin: Epithelium</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disease_Has_Primary_Anatomic_Site: Prostate Gland</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Synonym with source data</th>
<th>Synonym with source data</th>
<th>Synonym with source data</th>
<th>Synonym with source data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenocarcinoma of Prostate</td>
<td>Adenocarcinoma of the Prostate</td>
<td>Prostate Adenocarcinoma</td>
<td></td>
</tr>
</tbody>
</table>
Semantic metadata example: Agent

<Agent>
  <name>Taxol</name>
  <nSCNumber>007</nSCNumber>
</Agent>
### Why do you need metadata?

<table>
<thead>
<tr>
<th>Class/Attribute</th>
<th>Example Object Data</th>
<th>CIA Metadata</th>
<th>NCI Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent</td>
<td></td>
<td>A sworn intelligence agent; a spy</td>
<td>Chemical compound administered to a human being to treat a disease or condition, or prevent the onset of a disease or condition</td>
</tr>
<tr>
<td>Agent nSCNumber</td>
<td>007</td>
<td>Identifier given to an intelligence agent by the National Security Council</td>
<td>Identifier given to chemical compound by the US Food and Drug Administration Nomenclature Standards Committee</td>
</tr>
<tr>
<td>Agent name</td>
<td>Taxol</td>
<td>CIA code name given to intelligence agents</td>
<td>Common name of chemical compound used as an agent</td>
</tr>
</tbody>
</table>
Computable Interoperability

<table>
<thead>
<tr>
<th>Agent</th>
<th>C1708</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td></td>
</tr>
<tr>
<td>nSCNumber</td>
<td>C1708:41243</td>
</tr>
<tr>
<td>CTEPName</td>
<td></td>
</tr>
<tr>
<td>FDAIndID</td>
<td></td>
</tr>
<tr>
<td>IUPACName</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug</th>
<th>C1708</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td></td>
</tr>
<tr>
<td>NDCCode</td>
<td></td>
</tr>
<tr>
<td>approvalDate</td>
<td></td>
</tr>
<tr>
<td>approver</td>
<td></td>
</tr>
<tr>
<td>fdaCode</td>
<td>C1708:41243</td>
</tr>
</tbody>
</table>

My model

Your model
Tying it all together: The caCORE semantic management framework

---

CDEs

<table>
<thead>
<tr>
<th>Concept Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2223333</td>
</tr>
<tr>
<td>2223866</td>
</tr>
<tr>
<td>2223869</td>
</tr>
<tr>
<td>2223870</td>
</tr>
<tr>
<td>2223871</td>
</tr>
</tbody>
</table>

Desc. Logic

Concept Codes

---

Bioinformatics Objects

Common Data Elements

Enterprise Vocabulary
Cancer Data Standards Repository

- ISO/IEC 11179 Registry for Common Data Elements – units of semantic metadata
- Client for Enterprise Vocabulary: metadata constructed from controlled terminology and annotated with concept codes
- Precise specification of Classes, Attributes, Data Types, Permissible Values: **Strong typing** of data objects.
- Tools:
  - **UML Loader**: automatically register UML models as metadata components
  - **CDE Curation**: Fine tune metadata and constrain permissible values with data standards
  - **Form Builder**: Create standards-based data collection forms
  - **CDE Browser**: search and export metadata components
Common Security Module
caCORE Architecture

**Clients**
- HTTP Clients
- SOAP Clients
- Perl Clients
- Java Applications

**Middleware**
- Web Application Server
  - Interfaces
    - Java
    - SOAP
    - XML
  - Domain Objects [Gene, Disease, Agent, etc.]
  - Data Access Objects

**Data**
- Biomedical Data
- Common Data Elements
- Enterprise Vocabulary
- Authorization
Development and Deployment

Use Cases
Design
Test Plans
Iterative Development
Modeling
Unit Testing
User Guides
System Testing
Staging
Packaging
Production

DEV | QA | STAGE | PROD
caCORE Software Development Kit
caCORE SDK Components

- **UML Modeling Tool** (any with XMI export)
- **Semantic Connector** (concept binding utility)
- **UML Loader**
- **Codegen** (middleware code generator)
- **Security Adaptor** (Common Security Module)

caCORE SDK Generates a caBIG Silver-Compliant System
Professional Documentation

caCORE
SOFTWARE DEVELOPMENT KIT 1.0.3

Programmer’s Guide

NATIONAL INSTITUTES OF HEALTH
NATIONAL CANCER INSTITUTE
Center for Bioinformatics
caBIG UML Models Completed and in the Works at Cancer Centers for Silver Systems

- caBIO: General bioinformatics
- caDSR: ISO11179 metadata
- EVS: Vocabulary
- caMOD: Cancer Models
- MAGE 1.2: Microarray data
- CSM: Security
- Common: Provenance, DBxrefs
- caTISSUE: Pathology reports.
- gridPIR: Protein Information
- mzXML: mass spec proteomics data
- scanFeatures: Proteomics
- AML: Proteomics
- statml: Statistical markup model
- CAP: College of American Pathologists protocols for Breast, Lung, Prostate
- GoMiner: Text mining tool for GO
- caTI SSUE: Tissue banking
- protLIMS: Laboratory Information Management System for proteomics
- BRIDG: Clinical Trials
From Silver to Gold:

caGrid
caBIG Use Cases

- **Advertisement**
  - *Service Provider* composes service metadata describing the service and publishes it to grid.

- **Discovery**
  - *Researcher* (or application developer) specifies search criteria describing a service of interest
  - The research submits the discovery request to a discovery service, which identifies a list of services matching the criteria, and returns the list.

- **Query and Invocation**
  - *Researcher* (or application developer) instantiates the grid service and access its resources

- **Security**
  - *Service Provider* restricts access to service based upon authentication and authorization rules
caGrid Service-Oriented Architecture

OGSA Compliant - Service Oriented Architecture

- Service Registry
- ID Resolution
- Service Description
- Grid Communication Protocol
- Transport
- Workflow
- Functions

- BPEL
- Globus Toolkit
- OGSA-DAl
- GRAM
- Globus

- Schema Management
- Metadata Management
- Service Registry
- Functions Management
- Resource Management

- Mobius
- caCORE
- Globus
- GSI
- CAS
- myProxy
- Globus
Service Data Elements (SDEs) describe services so clients can discover what they do.

Two types of top-level grid services defined:
- Data Services
- Analytical Services

Three models for SDEs have been designed:
- Data service-specific
- Analytical Service-specific
- Common (all services)
Silver to Gold: Data Services

caBIG Gold data service

caGrid Infrastructure

Query Adaptor

Silver Data Service
Client and service APIs are object oriented, and operate over well-defined and curated data types.

Objects are defined in UML and converted into ISO/IEC 11179 Administered Components, which are in turn registered in the Cancer Data Standards Repository (caDSR).

Object definitions draw from vocabulary registered in the Enterprise Vocabulary Services (EVS), and their relationships are thus semantically described.

XML serialization of objects adhere to XML schemas registered in the Global Model Exchange (GME).
Analytical Services

- Accept and emit strongly typed data objects that conform to Gold data service requirements
- Analytical method implementation is defined by service provider
- Toolkit to assist with creating a caGrid Analytical Service will come with caGrid 0.5 download
Analytical Service Creation Wizard
Method Implementation

```java
package gov.nih.nci.cagrid.service;

import gov.nih.nci.cagrid.common.CaGRIDExampleI;
import org.globus.ogsa.GridServiceBase;

/**
 * CaGRIDExampleI TODO: DOCUMENT ME
 *
 * @created by CaGRID toolkit 0.5
 */
public class CaGRIDExampleImpl implements CaGRIDExampleI {

    private GridServiceBase base;

    public CaGRIDExampleImpl(GridServiceBase base) {
        this.base = base;
    }

    public gov.nih.nci.cagrid.bean.PersonType changeAddress(gov.nih.nci.cagrid.bean.PersonType input,
                                                            gov.nih.nci.cagrid.bean.AddressType address) {
        //TODO: Implement this autogenerated method
        return null;
    }
}
```

Insert method code here
caGrid 0.5 Test Bed
Acknowledgements

NCI
Andrew von Eschenbach
Anna Barker
Wendy Patterson
OC
DCTD
DCB
DCP
DCEG
DCCPS
CCR

NCI CB
Ken Buetow
Avinash Shanbhag
George Komatsoulis
Denise Warzel
Frank Hartel
Sherri De Coronado
Dianne Reeves
Gilberto Fragoso
Jill Hadfield
Sue Dubman
Leslie Derr

Industry Partners
SAIC
BAH
Oracle
ScenPro
Ekagra
Apelon
Terrapin Systems
Panther Informatics
Acknowledgements – caGrid

Georgetown
- Baris Suzek
- Scott Shung
- Colin Freas
- Nick Marcou
- Arnie Miles
- Cathy Wu
- Robert Clarke

NCICB
- Avinash Shanbhag
- George Komatsoulis
- Denise Warzel
- Frank Hartel

Duke
- Patrick McConnell

UPMC
- Rebecca Crawley
- Kevin Mitchell

TerpSys
- Gavin Brennan
- Troy Smith
- Wei Lu
- Doug Kanoza

Ohio State Univ.
- Scott Oster
- Shannon Hastings
- Steve Langella
- Tahsin Kurc
- Joel Saltz

SAIC
- William Sanchez
- Manav Kher
- Rouwei Wu
- Jijin Yan
- Tara Akhavan

Panther Informatics
- Brian Gilman
- Nick Encina

Oracle
- Ram Chilukuri

BAH
- Arumani Manisundaram
caBIG Participant Community

9Star Research
Albert Einstein
Ardais
Argonne National Laboratory
Burnham Institute
California Institute of Technology-JPL
City of Hope
Clinical Trial Information Service (CTIS)
Cold Spring Harbor
Columbia University-Herbert Irving
Consumer Advocates in Research and Related Activities (CARRA)
Dartmouth-Norris Cotton
Data Works Development
Department of Veterans Affairs
Drexel University
Duke University
EMMES Corporation
First Genetic Trust
Food and Drug Administration
Fox Chase
Fred Hutchinson
GE Global Research Center
Georgetown University-Lombardi
IBM
Indiana University
Internet 2
Jackson Laboratory
Johns Hopkins-Sidney Kimmel
Lawrence Berkeley National Laboratory
Massachusetts Institute of Technology
Mayo Clinic
Memorial Sloan Kettering
Meyer L. Prentis-Karmanos
New York University
Northwestern University-Robert H. Lurie
Ohio State University-Arthur G. James/Richard Solove
Oregon Health and Science University
Roswell Park Cancer Institute
St Jude Children's Research Hospital
Thomas Jefferson University-Kimmel
Translational Genomics Research Institute
Tulane University School of Medicine
University of Alabama at Birmingham
University of Arizona
University of California Irvine-Chao Family
University of California, San Francisco
University of California-Davis
University of Chicago
University of Colorado
University of Hawaii
University of Iowa-Holden
University of Michigan
University of Minnesota
University of Nebraska
University of North Carolina-Lineberger
University of Pennsylvania-Abramson
University of Pittsburgh
University of South Florida-H. Lee Moffitt
University of Southern California-Dr.aaS
University of Vermont
University of Wisconsin
Vanderbilt University-Ingram
Velos
Virginia Commonwealth University-Massey
Virginia Tech
Wake Forest University
Washington University-Siteman
Wistar
Yale University
From Village to City