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

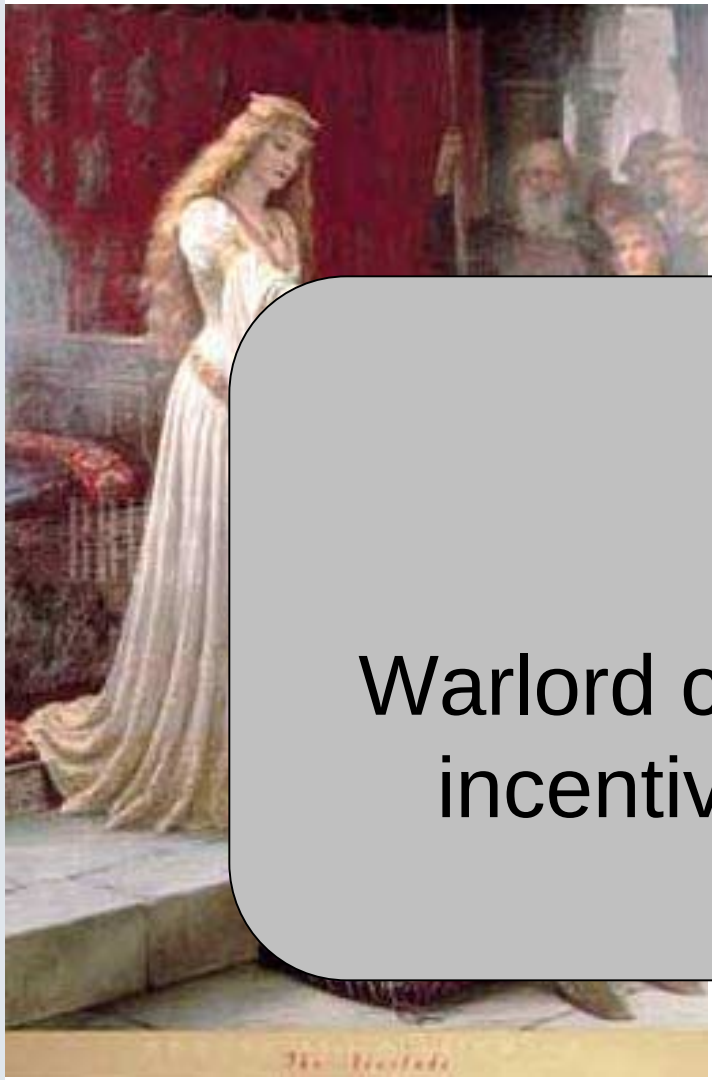
Scenario from caBIG Strategic Plan

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



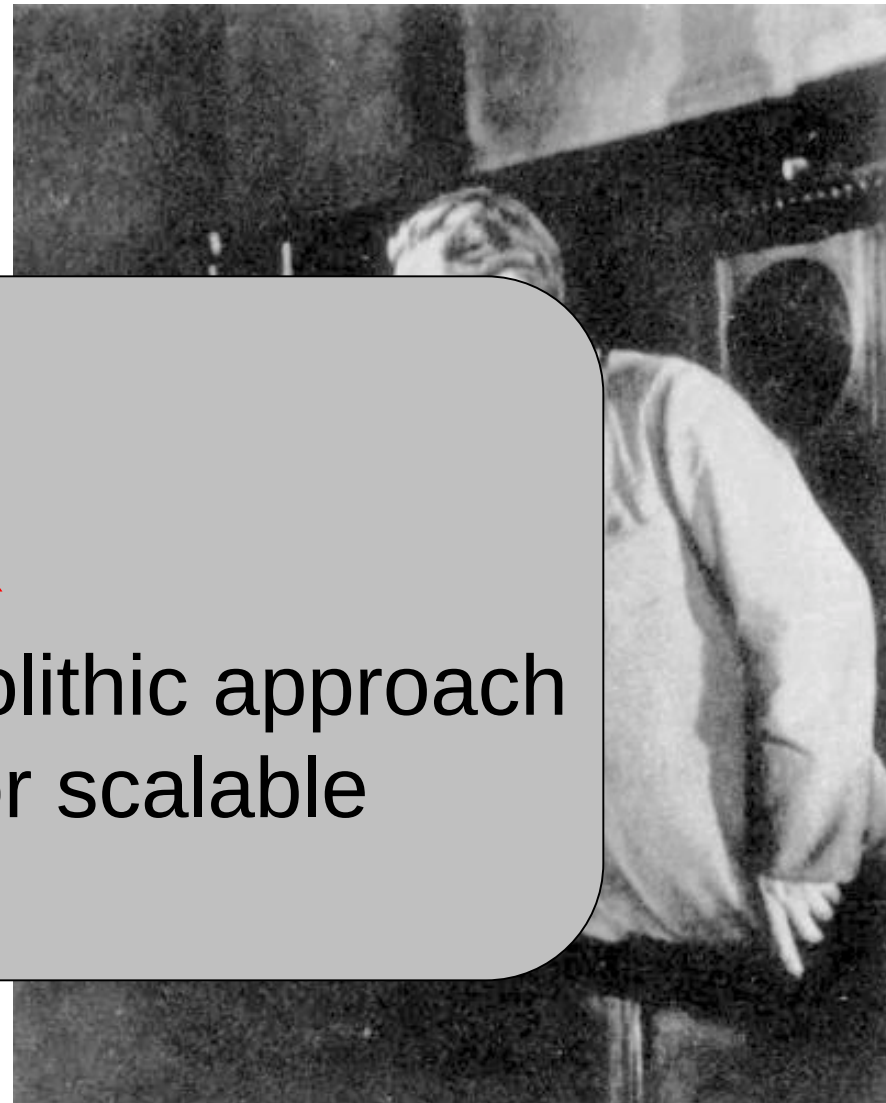
Feudalism

X

Warlord culture offers little incentive to cooperate

Governance Models

Forced Collectivization



X

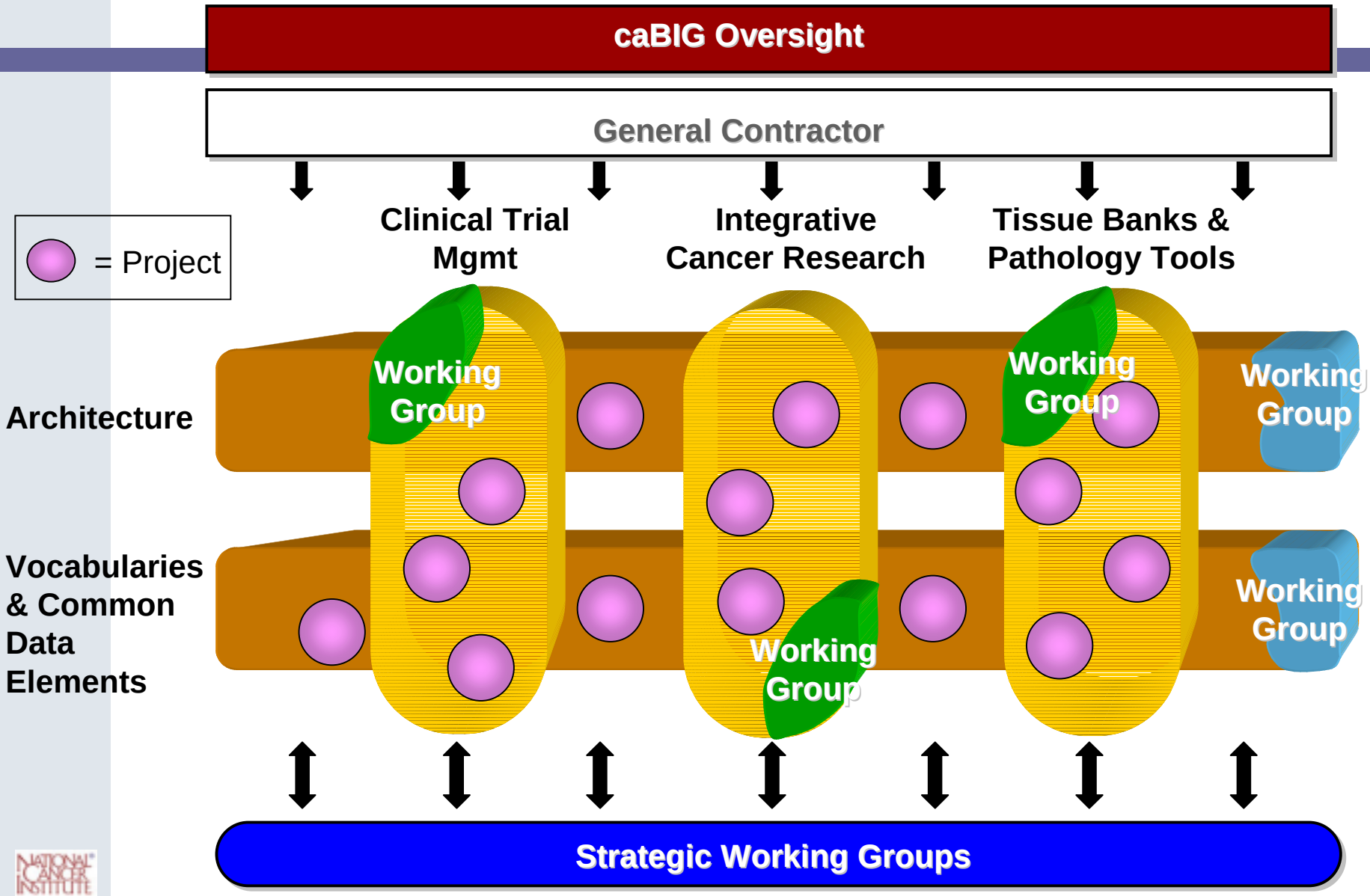
Centralized monolithic approach
not flexible or scalable

Governance Models

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

Federal Democracy

caBIG Organization Structure



Interoperability

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

Syntactic interoperability

Semantic interoperability

caBIG Compatibility Guidelines

SYNTACTIC

SEMANTIC

SEMANTIC

SEMANTIC

Maturity Model	Legacy	Bronze	Silver	Gold
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*	*- Use of publicly accessible controlled vocabularies as well as local terminologies.* *- All terminologies must include unambiguous definitions of terms*	*- 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.* *- Metadata is stored and publicized in an electronic format that is separate from the resource that is being described..*	*- 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*

NATIONAL CANCER INSTITUTE

Model-Driven Architecture

MDA - Microsoft Internet Explorer

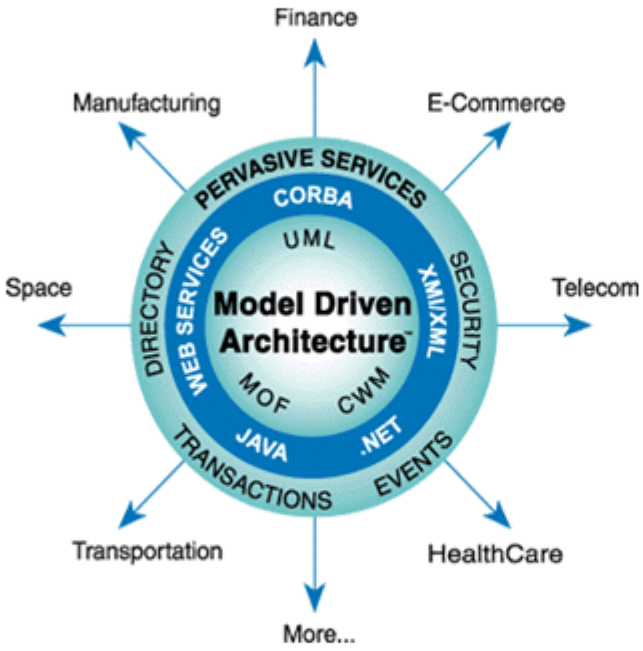
File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.omg.org/mda/ Go Links

- Exec Overview
- FAQ
- Presentations & Papers
- Specifications
- MDA Guide
- Education
- Industry Support
- Press Releases
- Committed Companies
- Success Stories
- FastStart Program
- News
- Reading Room
- Contacts

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).

Done Internet

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

caBIO

Bioinformatics Objects

caDSR

Common Data Elements

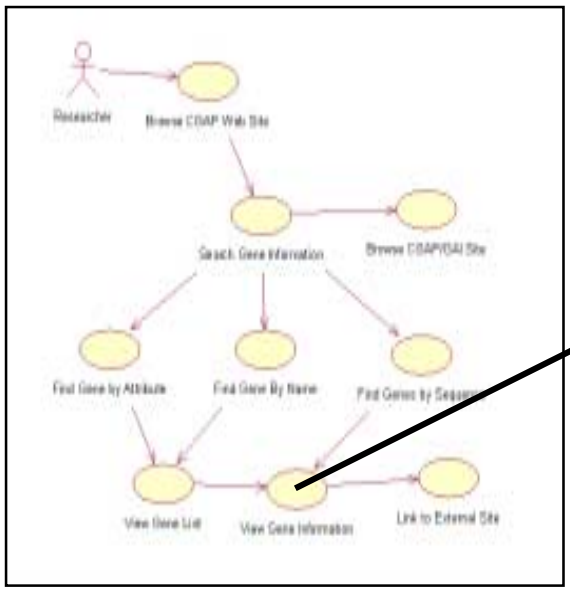
EVS

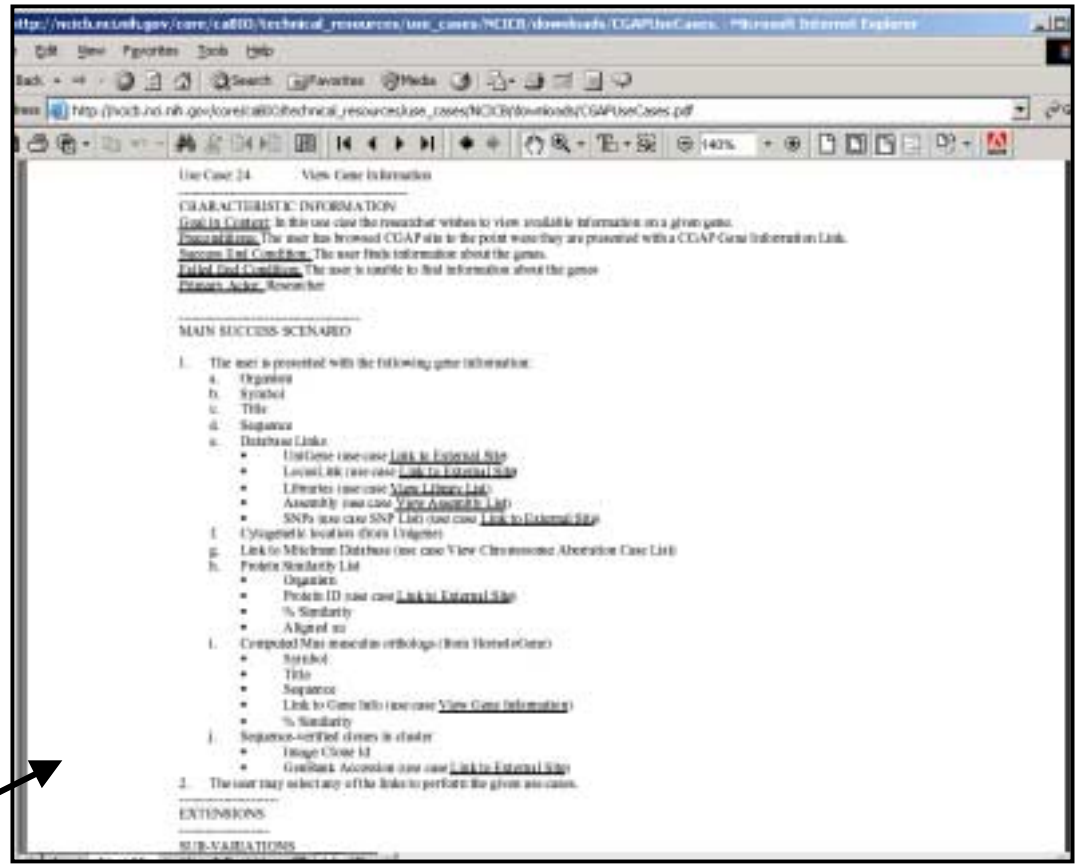
Enterprise Vocabulary

SECURITY

Use Cases

- Description
- Actors
- Basic Course
- Alternative Course





Use Case 24 View Gene Information

CHARACTERISTIC INFORMATION

Goal in Context: In this use case the researcher wishes to view available information on a given gene.

Preconditions: The user has browsed CGAP site to the point were they are presented with a CGAP Gene Information Link.

Success End Condition: The user finds information about the genes.

Failed End Condition: The user is unable to find information about the genes

Primary Actor: Researcher

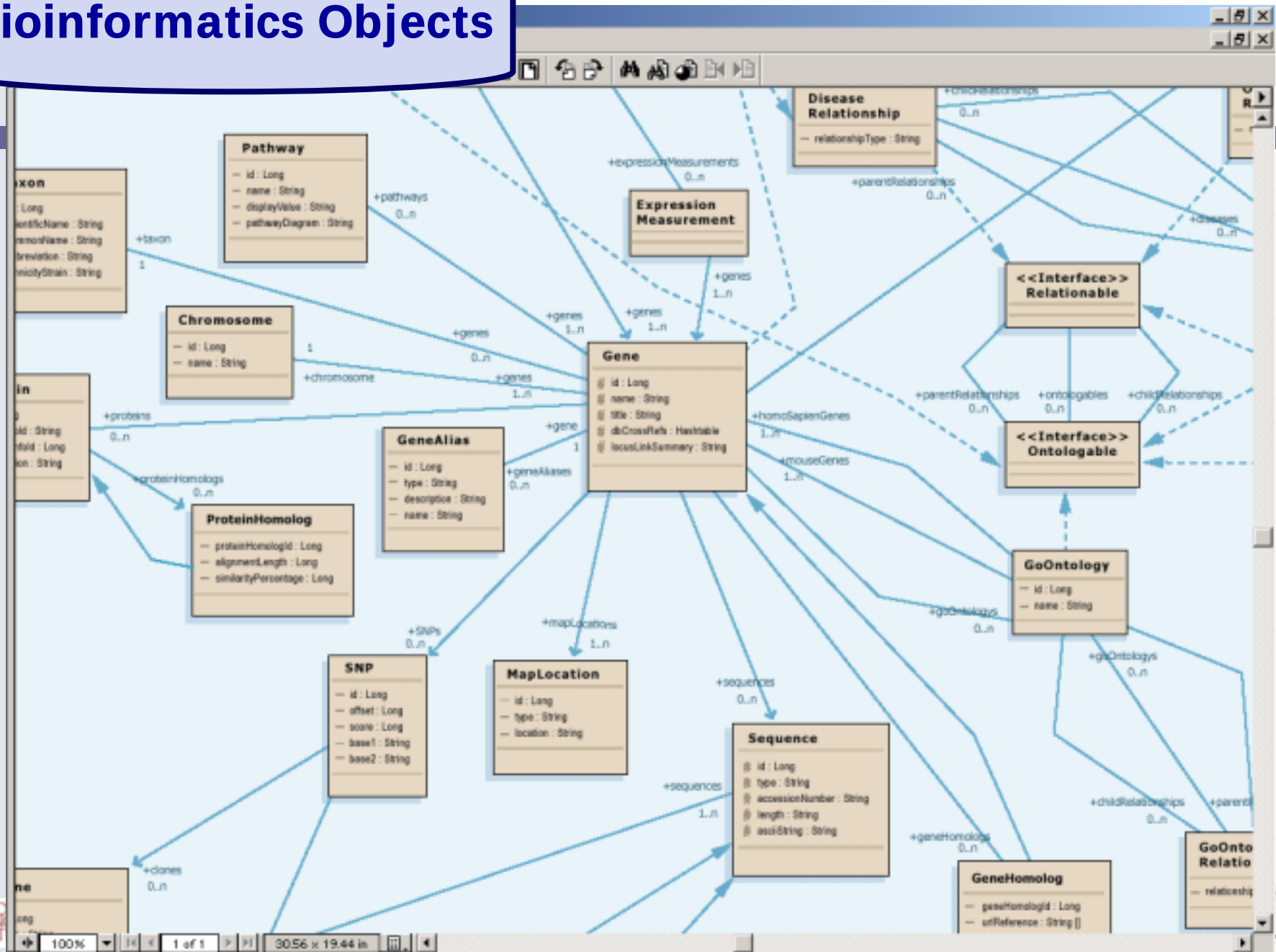
MAIN SUCCESS SCENARIO

1. The user is presented with the following gene information:
   a. Organism
   b. Symbol
   c. Title
   d. Sequence
   e. Database Links
      - UniGene (use case Link to External Site)
      - LocusLink (use case Link to External Site)
      - Libraries (use case View Library List)
      - Assembly (use case View Assembly List)
      - SNPs (use case SNP List) (use case Link to External Site)
   f. Cytogenetic location (from Unigene)
   g. Link to Mitelman Database (use case View Chromosome Aberration Case List)
   h. Protein Similarity List
      - Organism
      - Protein ID (use case Link to External Site)
      - % Similarity
      - Aligned to
   i. Computed Mus musculus orthologs (from HomoloGene)
      - Symbol
      - Title
      - Sequence
      - Link to Gene Info (use case View Gene Information)
      - % Similarity
   j. Sequence-verified clones in cluster
      - Image Clone Id
      - GenBank Accession (use case Link to External Site)
2. The user may select any of the links to perform the given use cases.

EXTENSIONS

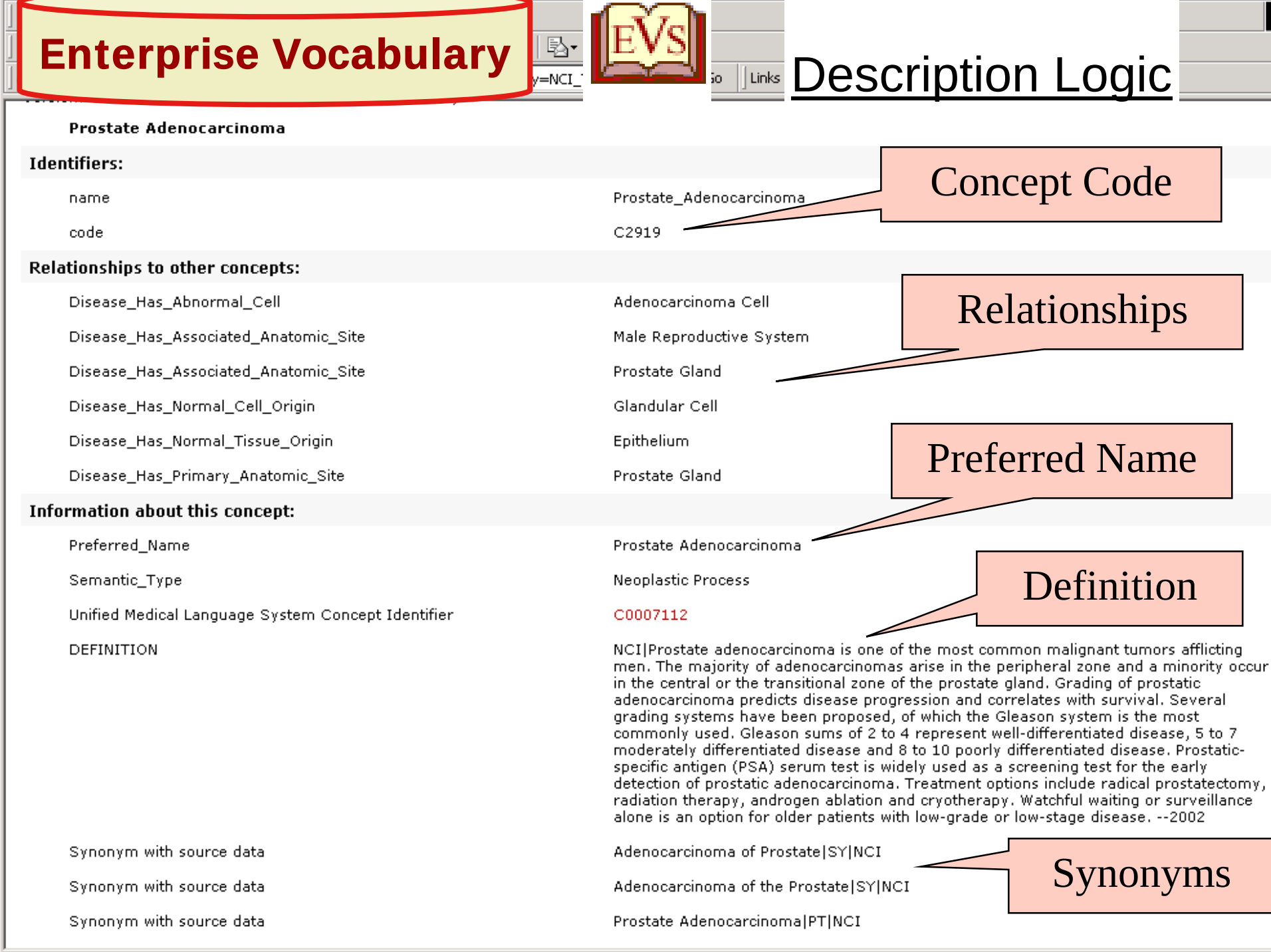
SUB-VARIATIONS

Bioinformatics Objects

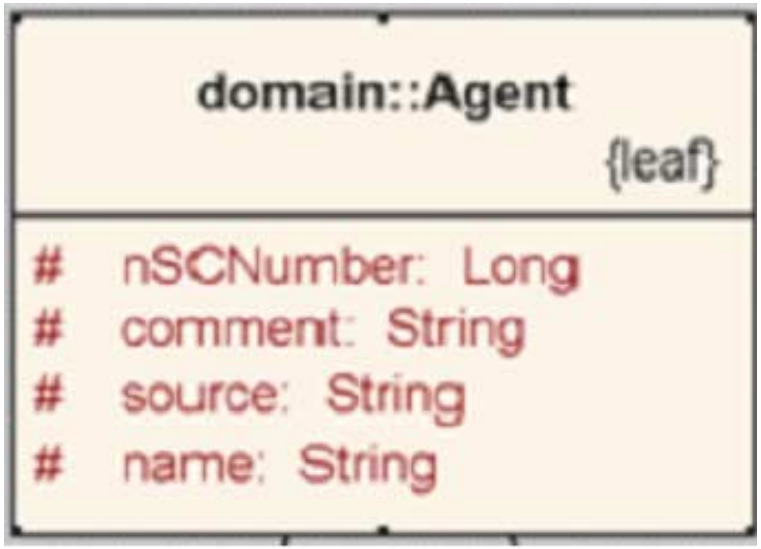


Common Data Elements

- 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


Enterprise Vocabulary
EVS
Description Logic
Links
Prostate Adenocarcinoma
Identifiers:
name
Prostate_Adenocarcinoma
code
C2919
Concept Code
Relationships to other concepts:
Disease_Has_Abnormal_Cell
Adenocarcinoma Cell
Disease_Has_Associated_Anatomic_Site
Male Reproductive System
Disease_Has_Associated_Anatomic_Site
Prostate Gland
Relationships
Disease_Has_Normal_Cell_Origin
Glandular Cell
Disease_Has_Normal_Tissue_Origin
Epithelium
Disease_Has_Primary_Anatomic_Site
Prostate Gland
Preferred Name
Information about this concept:
Preferred_Name
Prostate Adenocarcinoma
Semantic_Type
Neoplastic Process
Unified Medical Language System Concept Identifier
C0007112
Definition
DEFINITION
NCI|Prostate adenocarcinoma is one of the most common malignant tumors afflicting men. The majority of adenocarcinomas arise in the peripheral zone and a minority occur in the central or the transitional zone of the prostate gland. Grading of prostatic adenocarcinoma predicts disease progression and correlates with survival. Several grading systems have been proposed, of which the Gleason system is the most commonly used. Gleason sums of 2 to 4 represent well-differentiated disease, 5 to 7 moderately differentiated disease and 8 to 10 poorly differentiated disease. Prostatic-specific antigen (PSA) serum test is widely used as a screening test for the early detection of prostatic adenocarcinoma. Treatment options include radical prostatectomy, radiation therapy, androgen ablation and cryotherapy. Watchful waiting or surveillance alone is an option for older patients with low-grade or low-stage disease. --2002
Synonym with source data
Adenocarcinoma of Prostate|SY|NCI
Synonym with source data
Adenocarcinoma of the Prostate|SY|NCI
Synonym with source data
Prostate Adenocarcinoma|PT|NCI
Synonyms

Semantic metadata example: Agent

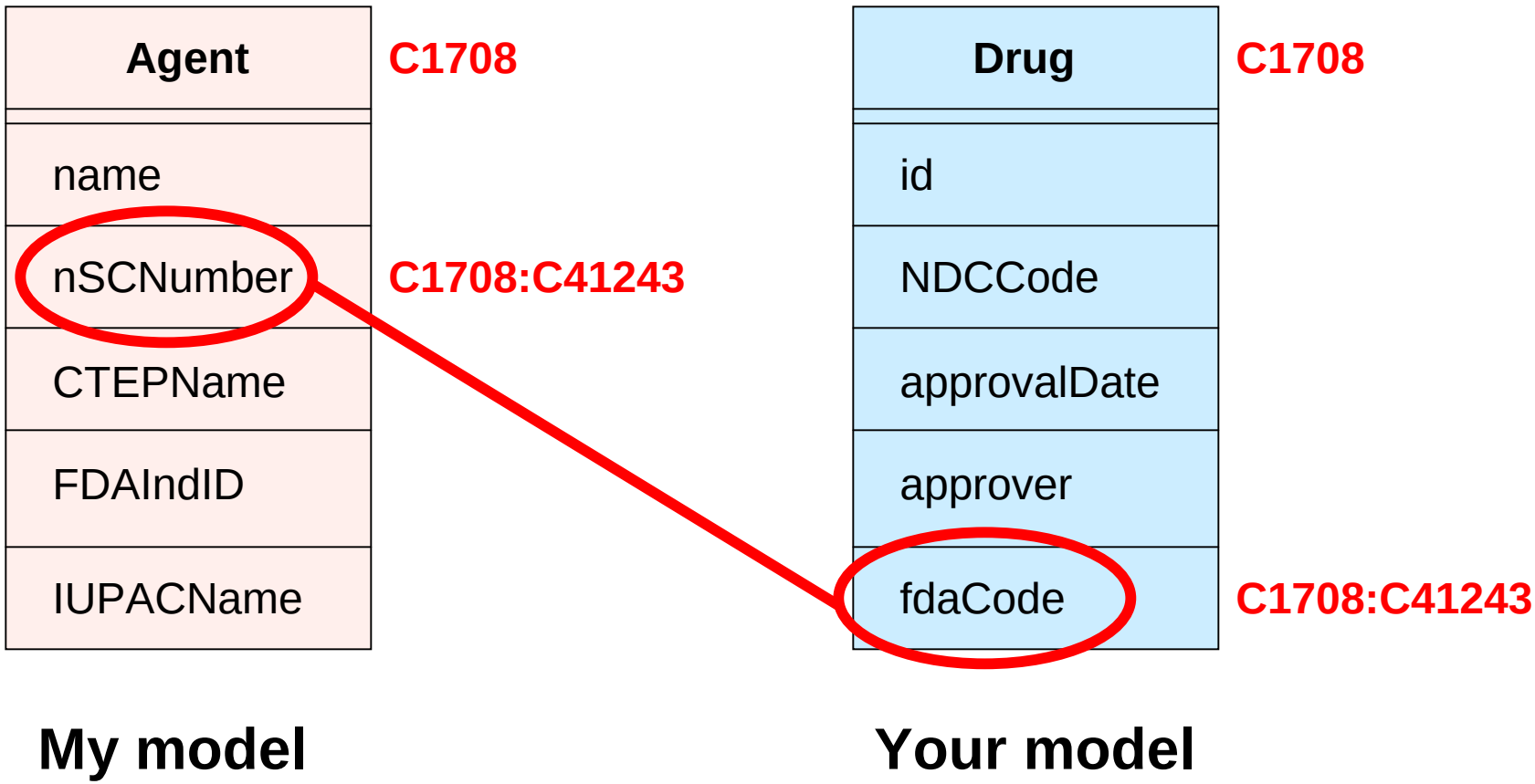



```
<Agent>
   <name>Taxol</name>
   <nSCNumber>007</nSCNumber>
</Agent>
```

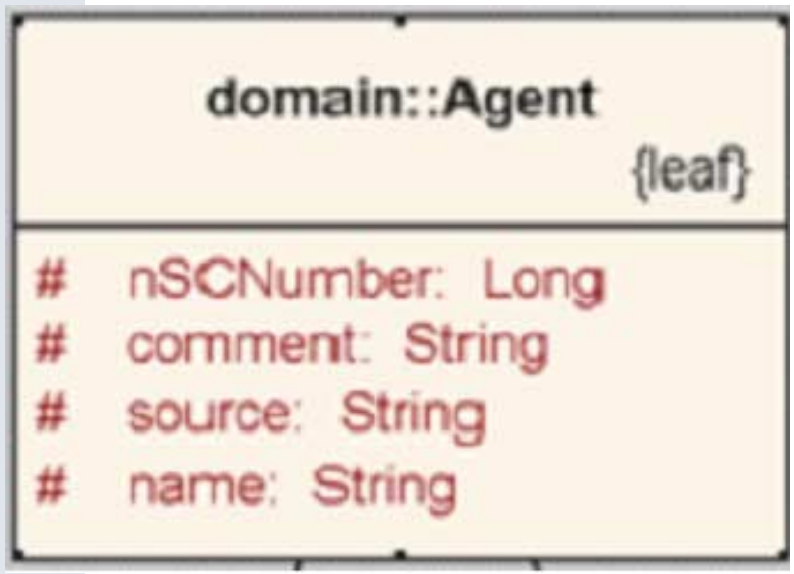
Why do you need metadata?

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

Computable Interoperability



Tying it all together: The caCORE semantic management framework



CDEs	Desc. Logic Concept Codes
2223333	C1708
2223866	C1708:C41243
2223869	C1708:C25393
2223870	C1708:C25683
2223871	C1708:C42614

Bioinformatics Objects

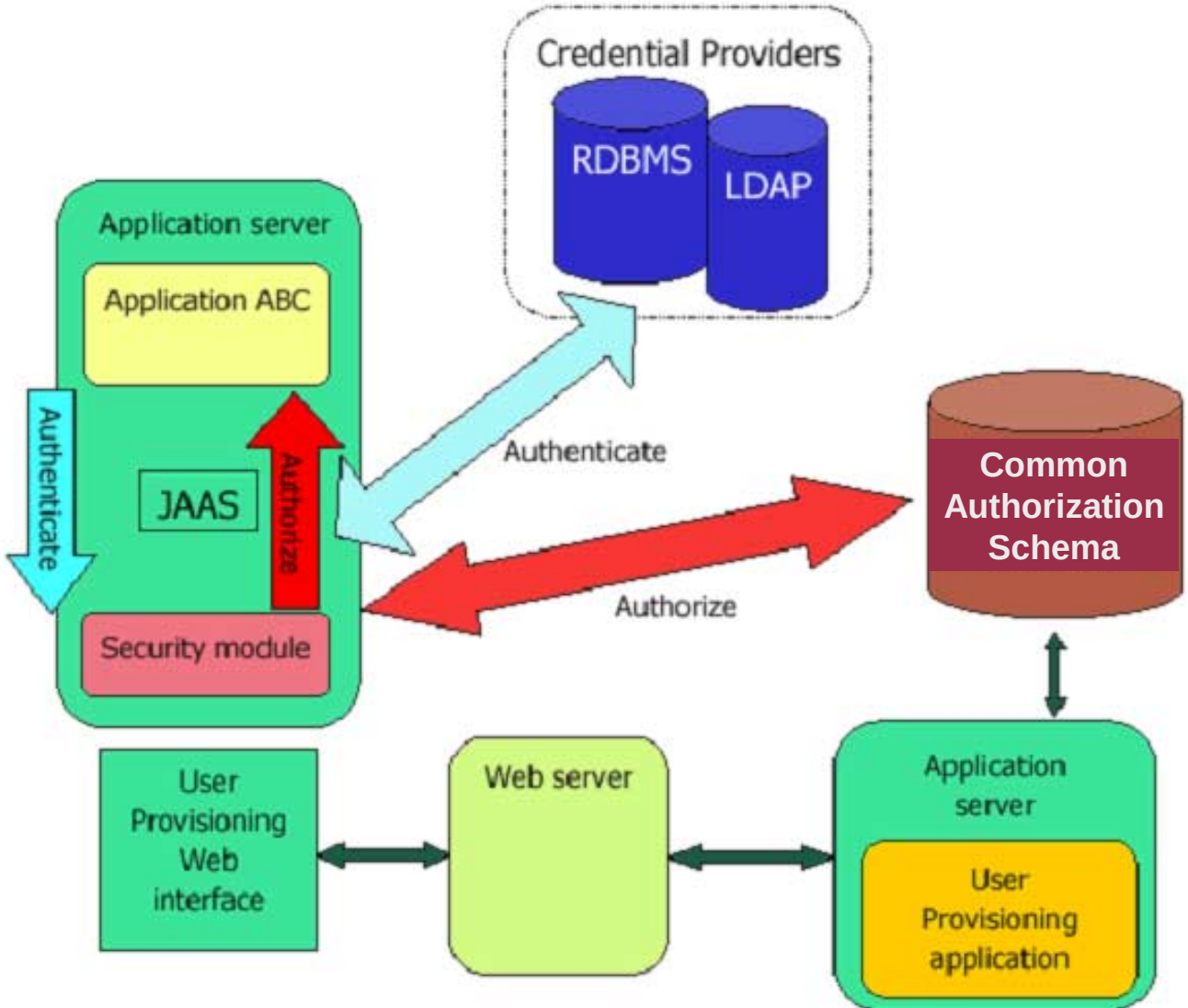
Common Data Elements

Enterprise Vocabulary

Cancer Data Standards Repository

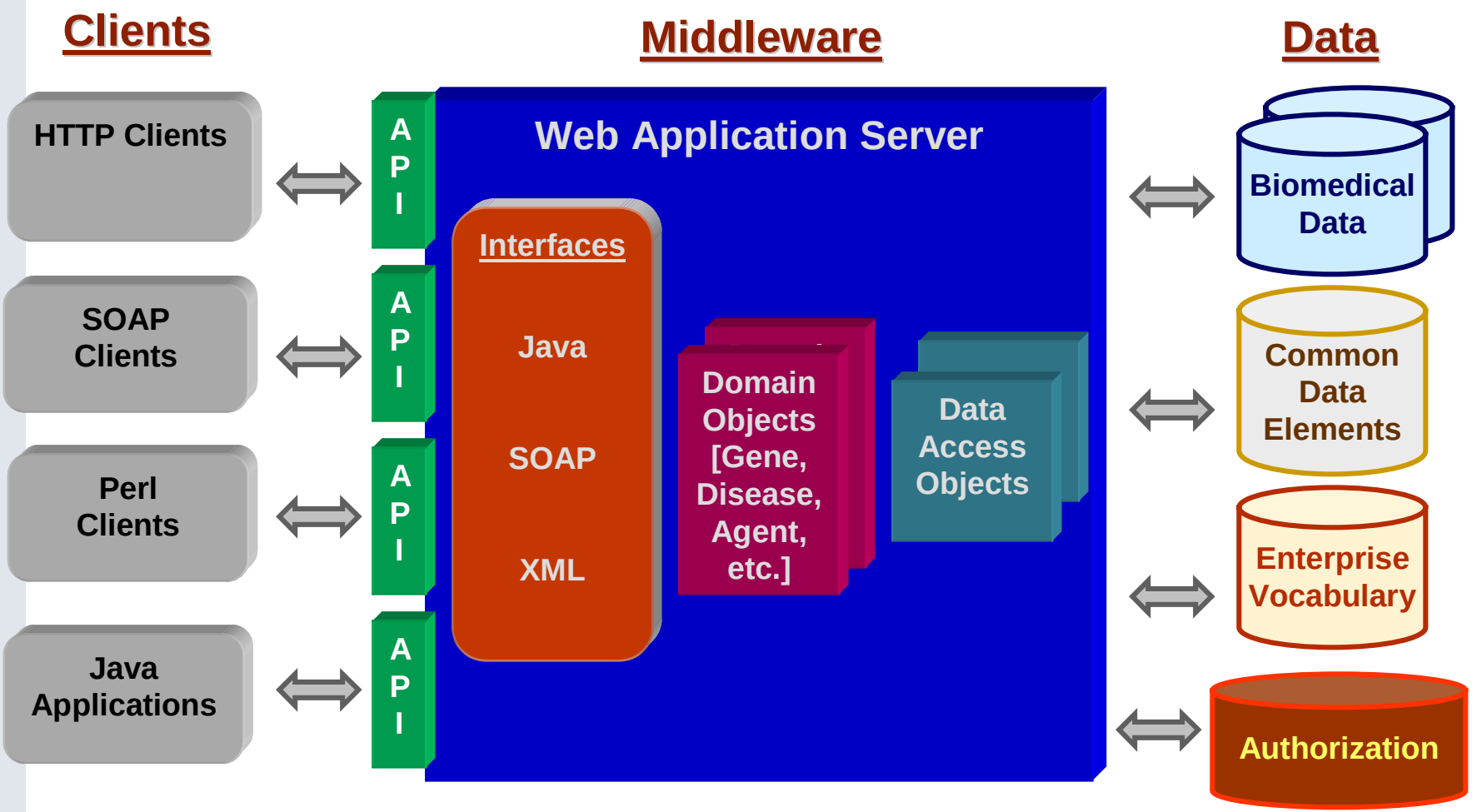
- ISO/IEC 11179 Registry for Common Data Elements – units of semantic metadata
- Client for Enterprise Vocabulary: metadata constructed from controlled terminology and annotated with concept codes
- Precise specification of Classes, Attributes, Data Types, Permissible Values: **Strong typing** of data 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

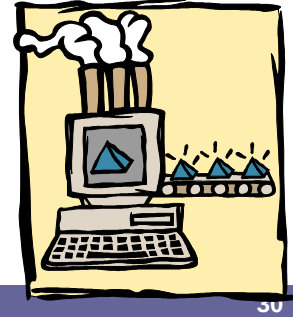


SECURITY

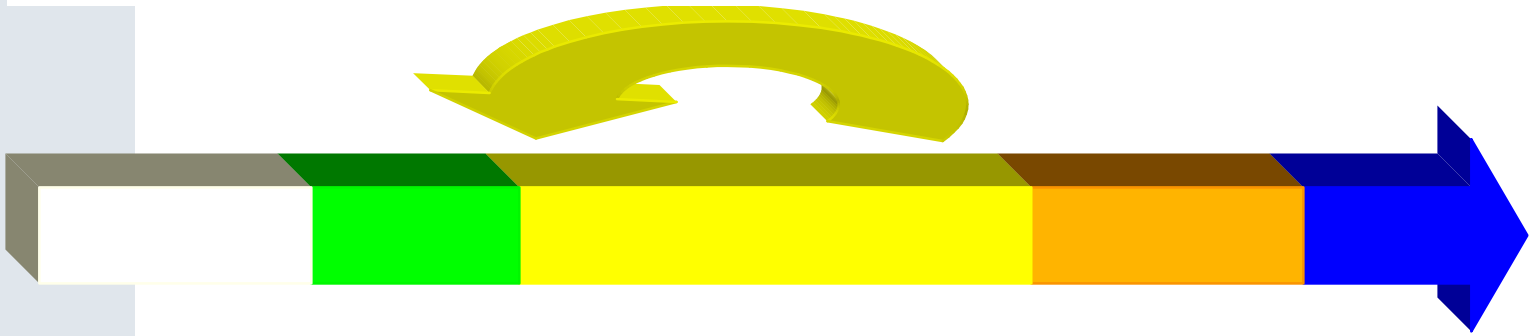
caCORE Architecture



Development and Deployment



DEV……..|QA……..|STAGE…|PROD



PRODUCTION

Use Cases

Design

Test Plans

Iterative Development

Modeling

Unit Testing

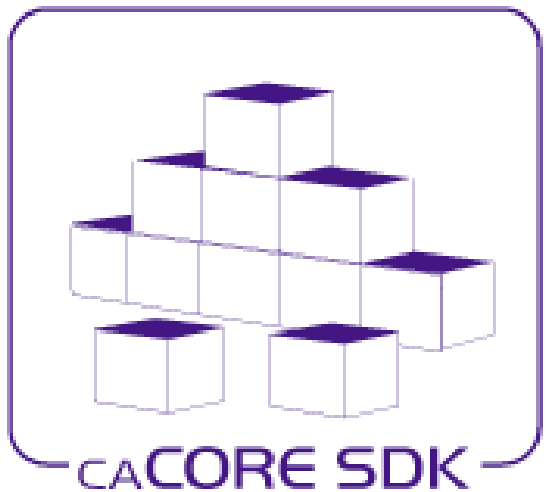
User Guides

System Testing

Staging

Packaging

caCORE Software Development Kit

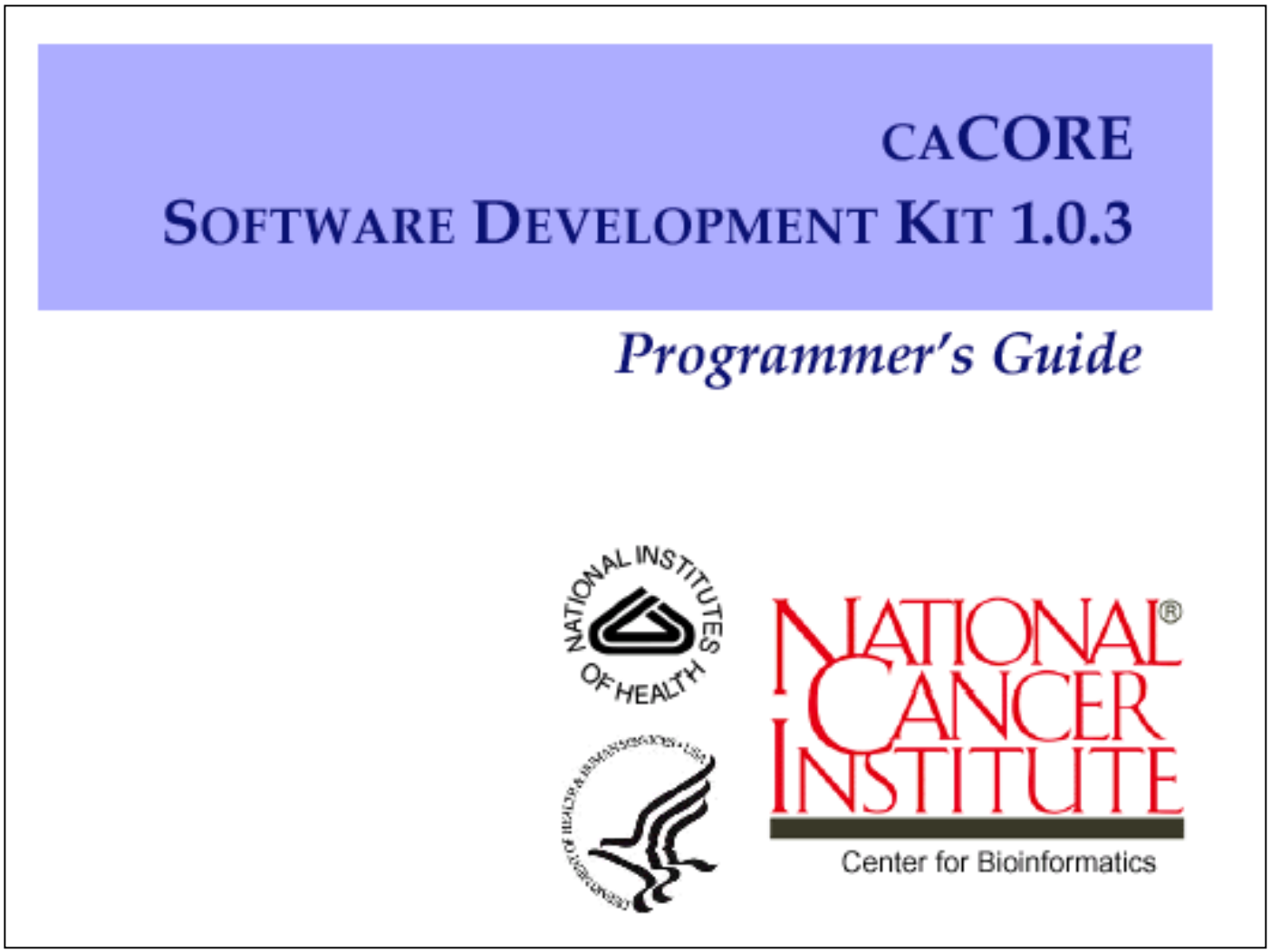


caCORE SDK Components

- **UML Modeling Tool** (any with XMI export)
- **Semantic Connector** (concept binding utility)
- **UI**
- **Co**
- **Se** ule)

caCORE SDK Generates a
caBIG Silver-Compliant System

Professional Documentation



caBIG UML Models Completed and in the Works at Cancer Centers for Silver Systems

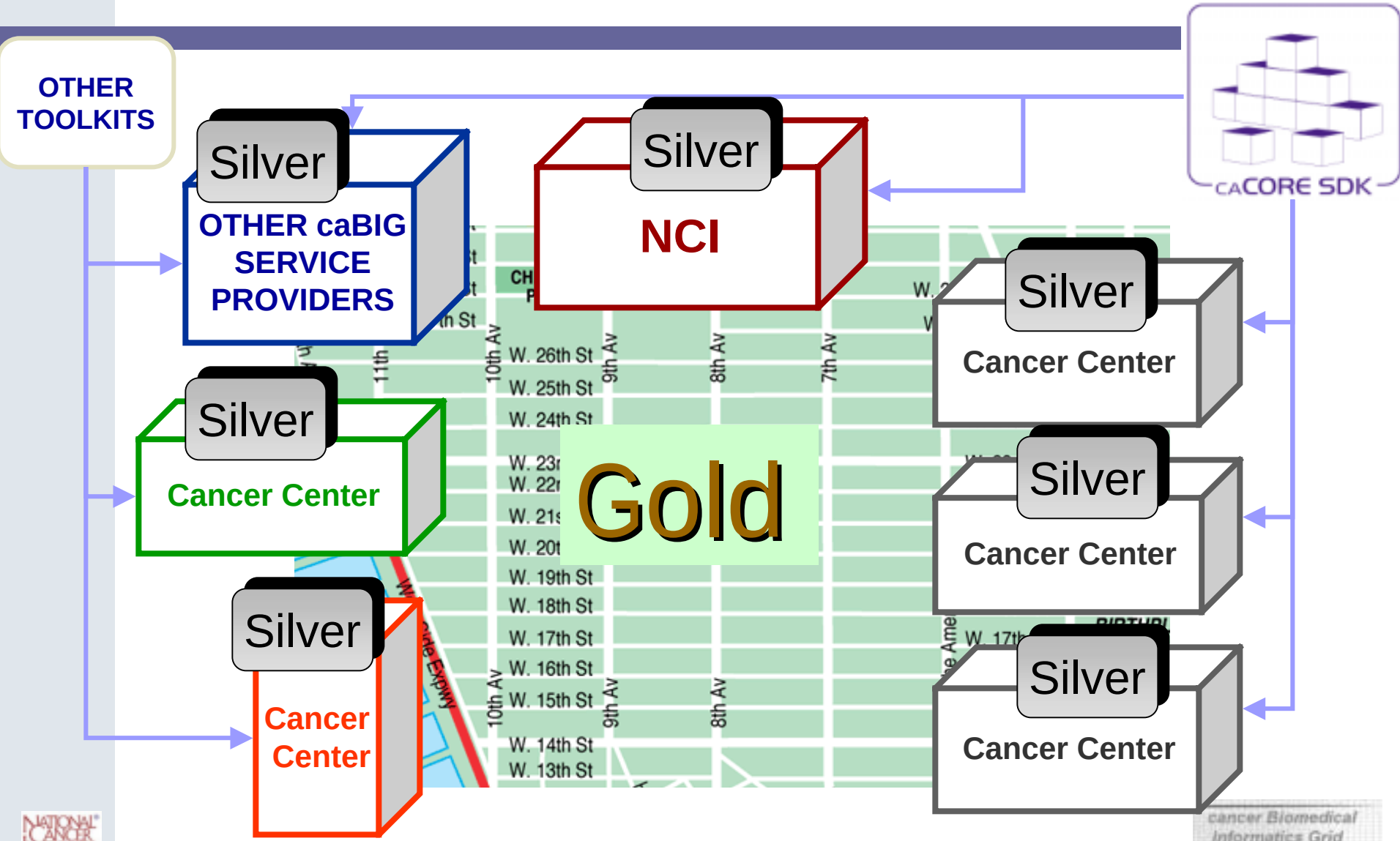
- caBIO General bioinformatics
- caDSR ISO11179 metadata
- EVS Vocabulary
- caMOD Cancer Models
- MAGE 1.2 Microarray data
- CSM Security
- Common Provenance, DBxrefs
- caTIES Pathology reports.
- gridPIR Protein Information
- mzXML mass spec proteomics data
  scanFeatures Proteomics
- AML Proteomics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

From Silver to Gold:

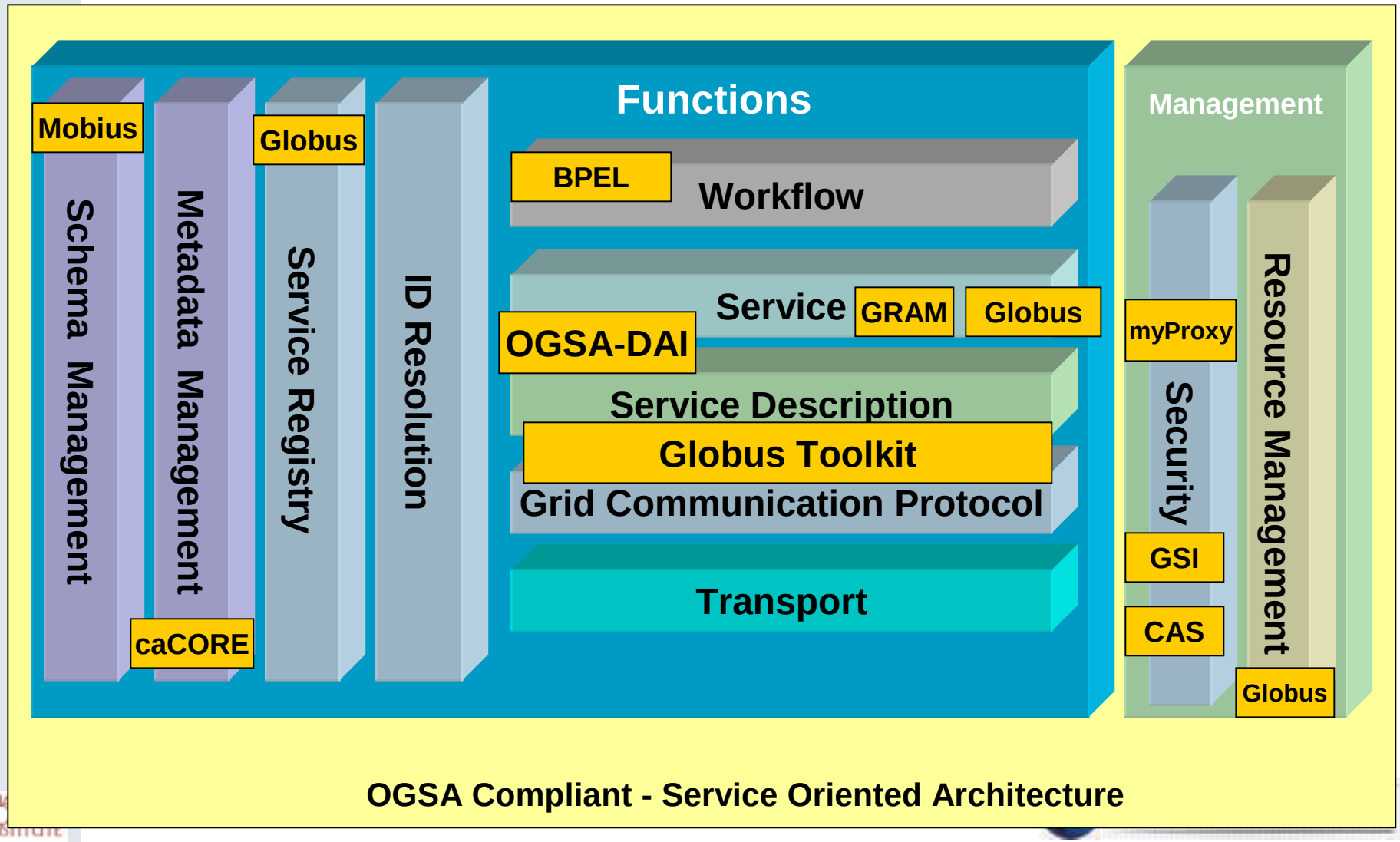
caGrid

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


OTHER TOOLKITS
caCORE SDK
Silver
OTHER caBIG SERVICE PROVIDERS
Silver
NCI
Silver
Cancer Center
Silver
Cancer Center
Gold
Silver
Cancer Center
Silver
Cancer Center
Silver
Cancer Center
W. 26th St
W. 25th St
W. 24th St
W. 19th St
W. 18th St
W. 17th St
W. 16th St
W. 15th St
W. 14th St
W. 13th St
10th Av
9th Av
8th Av
7th Av
11th

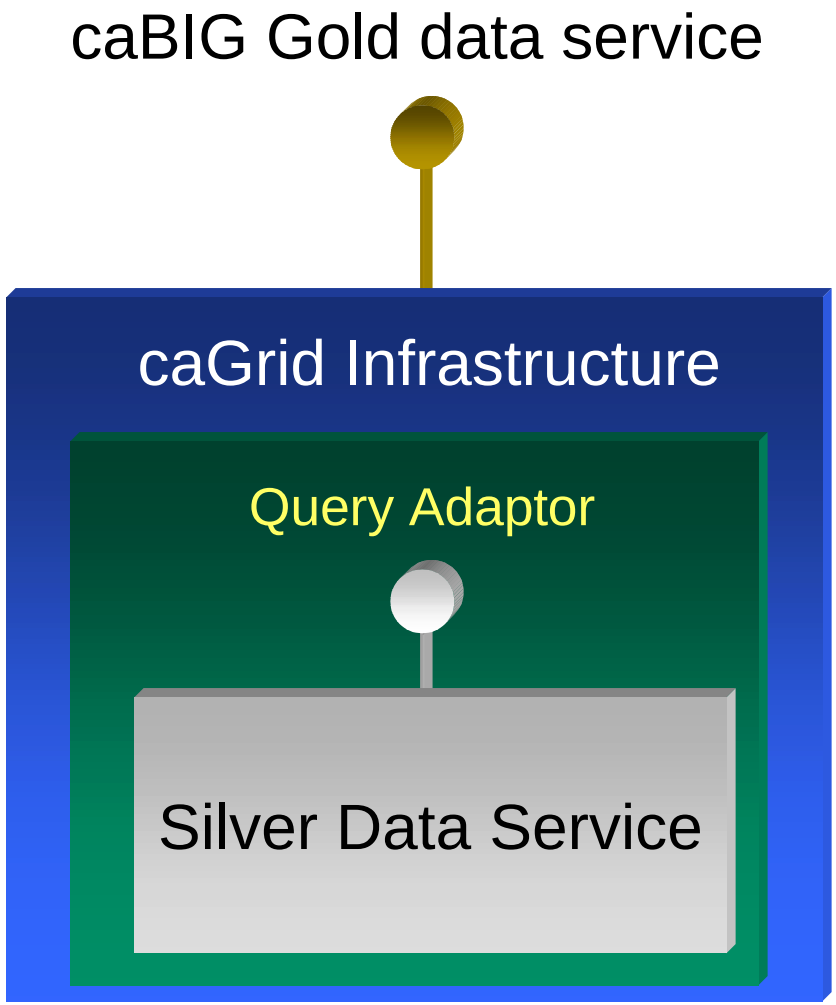
caGrid Service-Oriented Architecture



Service Data Elements

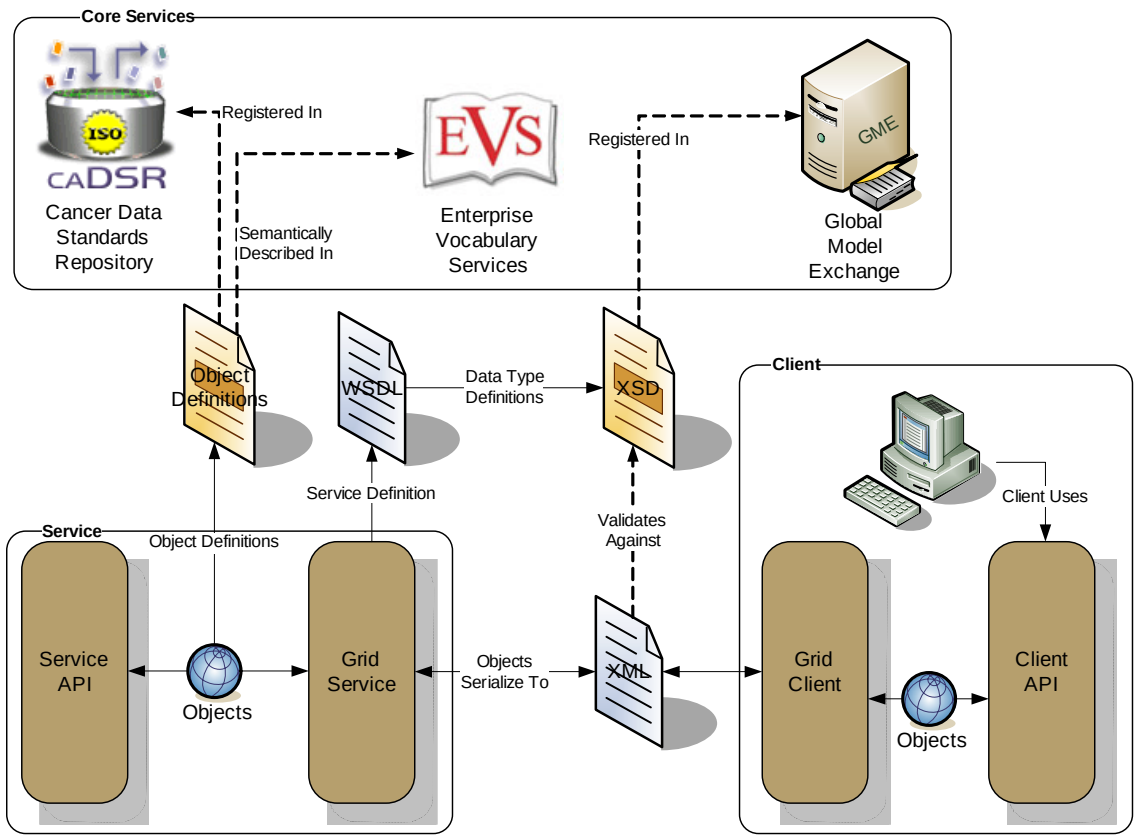
- Service Data Elements (SDEs) describe services so clients can discover what they do
- Two types of top-level grid services defined
  - **Data Services**
  - **Analytical Services**
- Three models for SDEs have been designed
  - Data service-specific
  - Analytical Service-specific
  - Common (all services)

Silver to Gold: Data Services



Data Object Semantics, Metadata, and Schemas

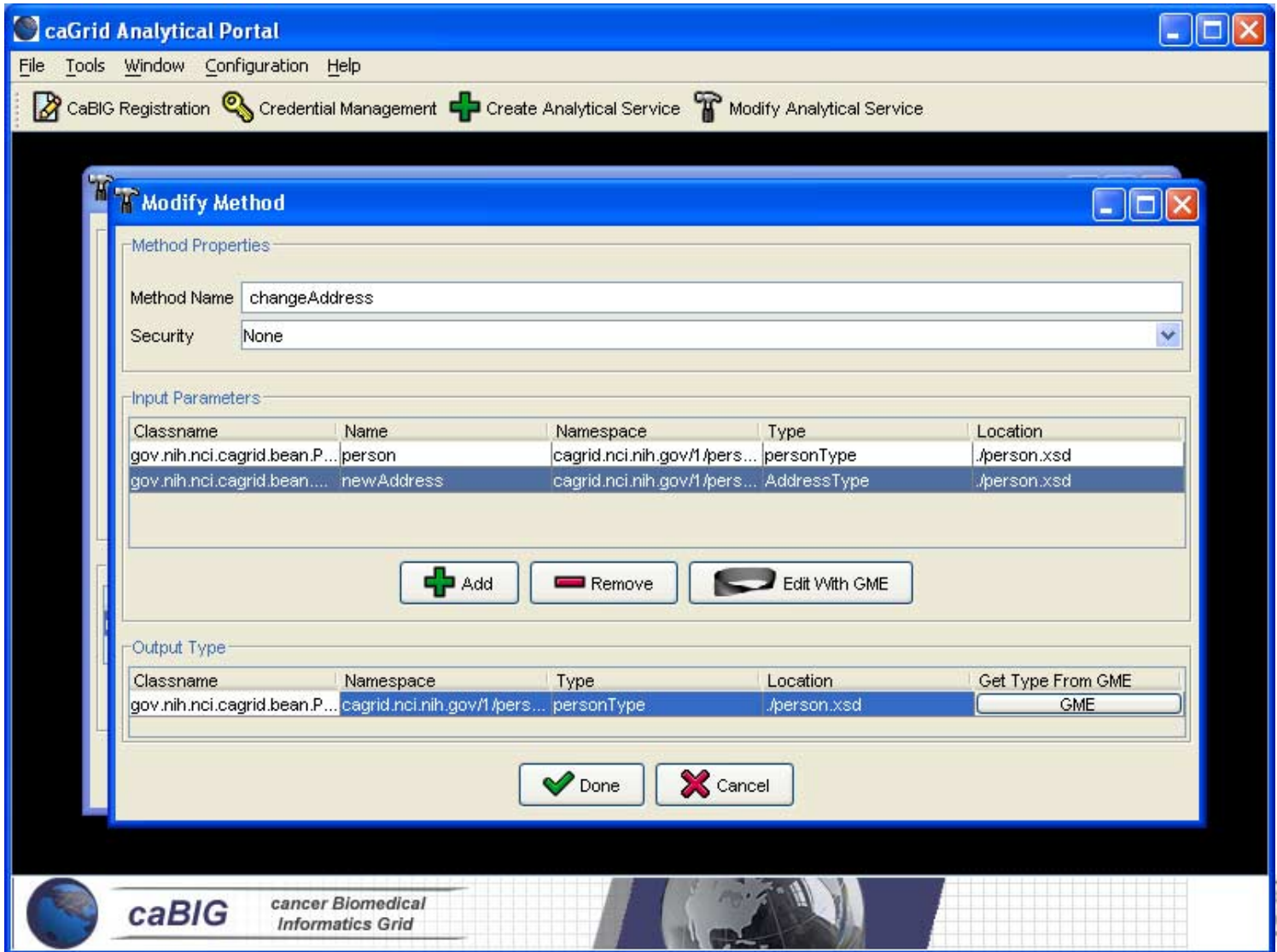
- Client and service APIs are object oriented, and operate over well-defined and curated data types
- Objects are defined in UML and converted into ISO/IEC 11179 Administered Components, which are in turn registered in the Cancer Data Standards Repository (caDSR)
- Object definitions draw from vocabulary 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

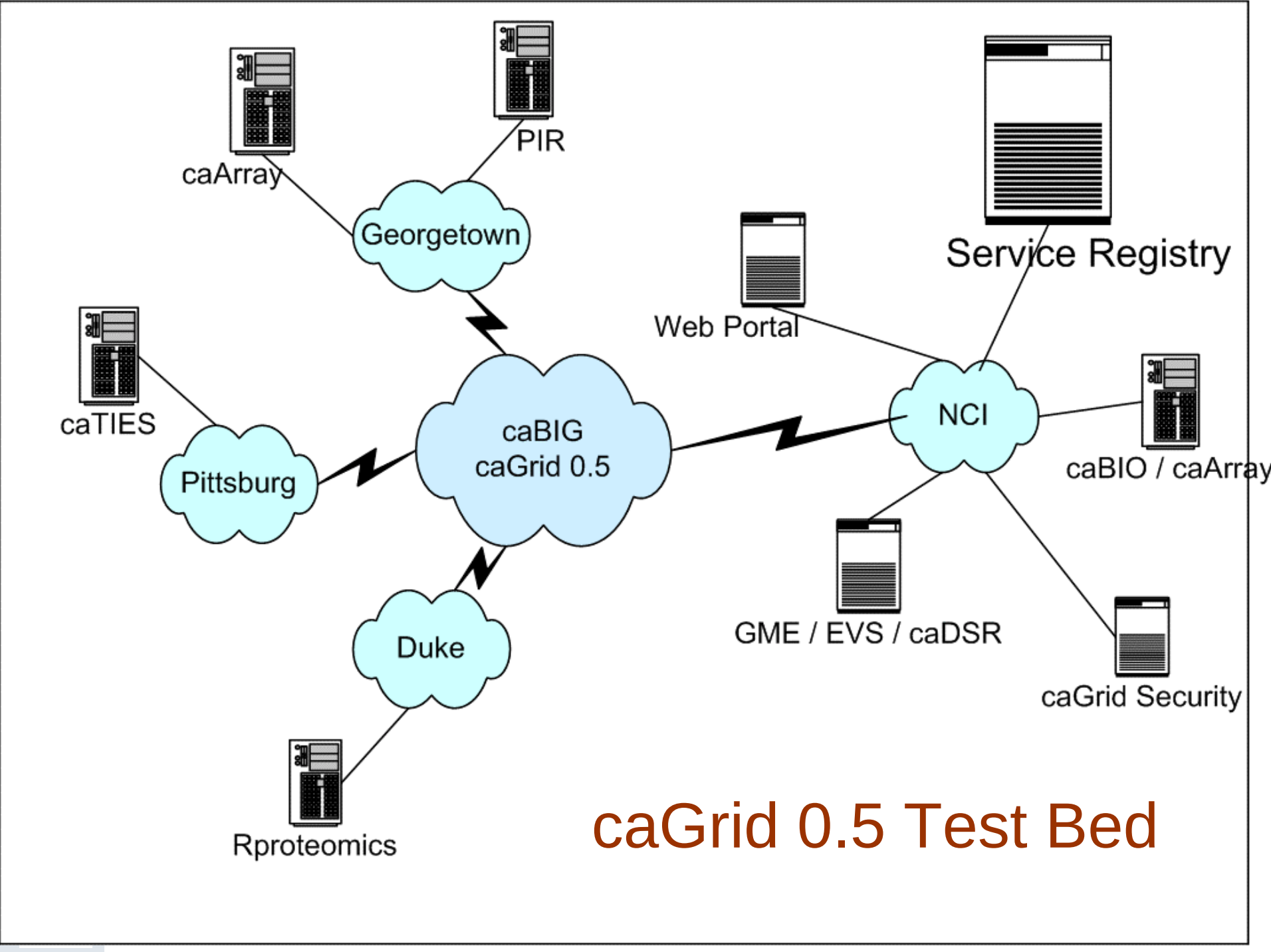
Analytical Service Creation Wizard



Method Implementation


```
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


caArray
PIR
Georgetown
Service Registry
Web Portal
caTIES
Pittsburg
caBIG
caGrid 0.5
NCI
caBIO / caArray
GME / EVS / caDSR
caGrid Security
Duke
Rproteomics
caGrid 0.5 Test Bed

Acknowledgements

NCI

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

NCICB

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
- Duke
  - Patrick McConnell
- UPMC
  - Rebecca Crawley
  - Kevin Mitchell
- NCICB
  - Avinash Shanbhag
  - George Komatsoulis
  - Denise Warzel
  - Frank Hartel
- 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-Norris
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