
# caCORE:

## A Common Framework for Creating, Managing and Deploying Semantically Interoperable Systems

SCIop

April 27, 2006

Denise Warzel

Associate Director, Core Infrastructure

National Cancer Institute

Center for Bioinformatics

# National Cancer Institute 2015 Goal

Relieve suffering and death due to cancer by the year 2015

# NCI Center for Bioinformatics (NCICB)

- 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.

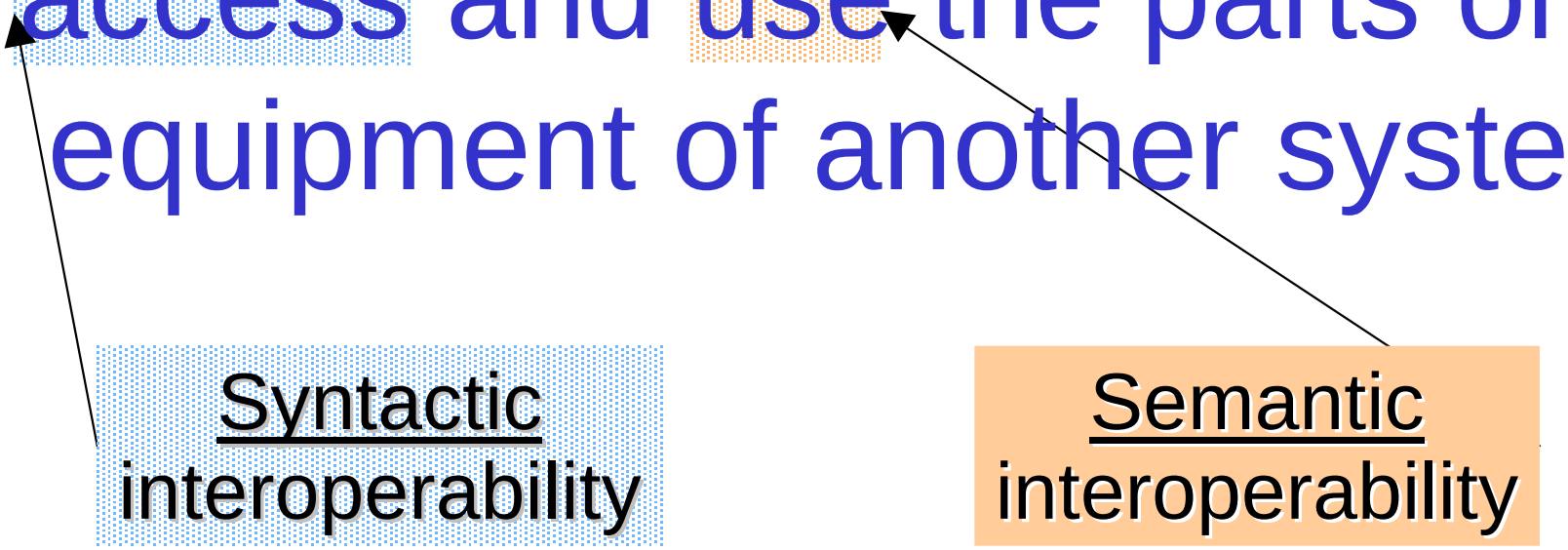
# NCICB Operations teams

- Systems and Hardware Support
- Database Administration
- Software Development
- Quality Assurance
- Technical Writing
- Application Support and Training
- caBIG Management

# Interoperability

ability of a system to

access and use the parts or equipment of another system



# Creating a Semantic Computing Infrastructure

- Issues to consider:
  - How will the standards get into the registry?
  - How will they be kept up to date managed throughout their Lifecycle?
  - How will the public access and use them?
  - How will software applications access and use them?

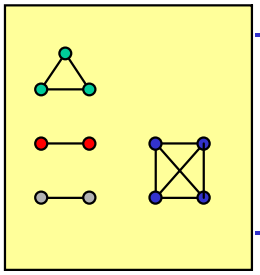
- NCI's approach: Build an infrastructure and tooling around the creation and management well formed, semantically unambiguous metadata
  - caCORE is the open-source foundation upon which the NCICB builds its data and information management systems

# Approaches: Semantic Integration and Interoperability

- Option 1
  - "Forced Collectivization"
  - Everyone adopts a single data model for a particular domain
  - Genbank, PDB, HL7 are examples of these sorts of models

  - Advantages:
    - Ensures interoperability
    - Minimal overhead
  - Disadvantages:
    - Not flexible
    - Does not allow for data stores for particular use cases

- Option 2:
  - "Local Networks"
  - Several sites agree on a format for interchanging data
  - Sites maintain a local data dictionary, XML schema, etc. to describe information model



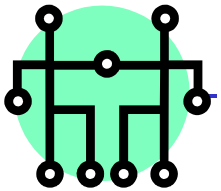
  - Advantages:
    - Flexible
    - Low Overhead
  - Disadvantages:
    - Works only where existing bilateral (or multilateral) agreements exist
    - Each new node must arrange to be interoperable with all other nodes or node cluster



** Derived from slides by G. Komatsoulis NCICB*

# Approaches: for Semantic Integration and Interoperability

- Option 3
  - "Common Data Elements"
  - Provide a complete description of all attributes in a systematic, uniform and unambiguous format
  - Description must be based on a common (but expandable) vocabulary.
  - Rely on concept codes, not concept names

  - Advantages:
    - Provides more ways to surface semantic matches – words and immutable codes
    - Allows new systems to find points of interoperability with all other data systems at once
    - Machine understandable
    - Stable immutable identifiers
  - Disadvantages:
    - Requires a very complete description of the data.
    - Some degree of overhead associated with creating and maintaining a compatible system

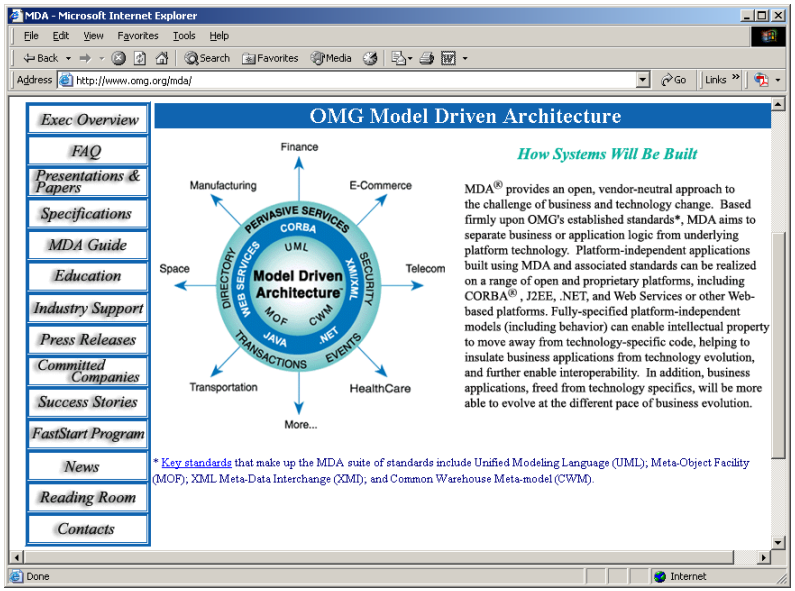
• ***Based on ISO 11179 Information Technology – Metadata Registries (MDR) parts 1-6***

# OMG MDA Approach

# Limitations of MDA

- 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

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



# caCORE – MDA plus a whole lot more!

caBIO

caDSR

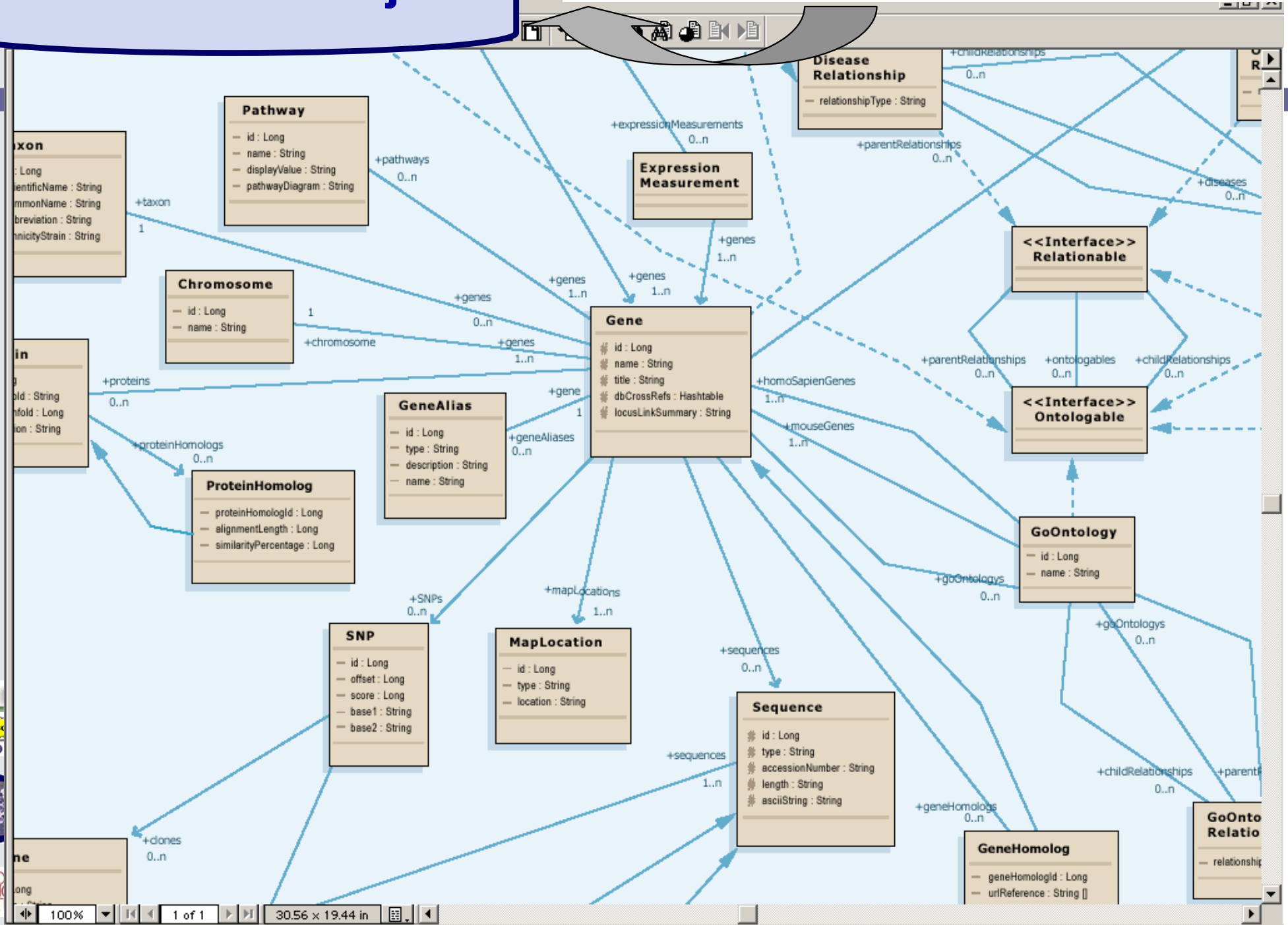
EVS


SECURITY

# Bioinformatics Objects

OMG OBJECT MANAGEMENT GROUP

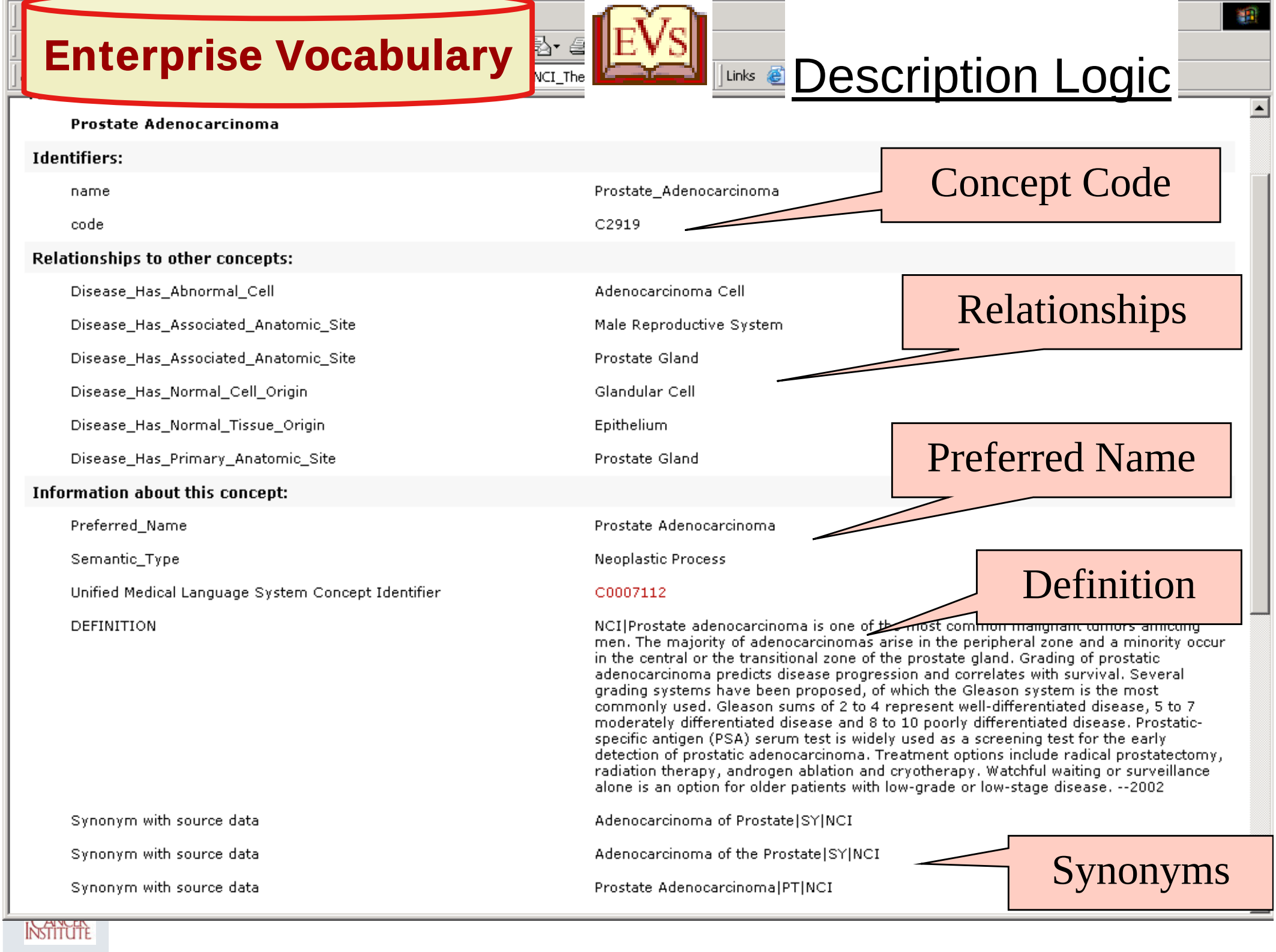
UNIFIED MODELING LANGUAGE™



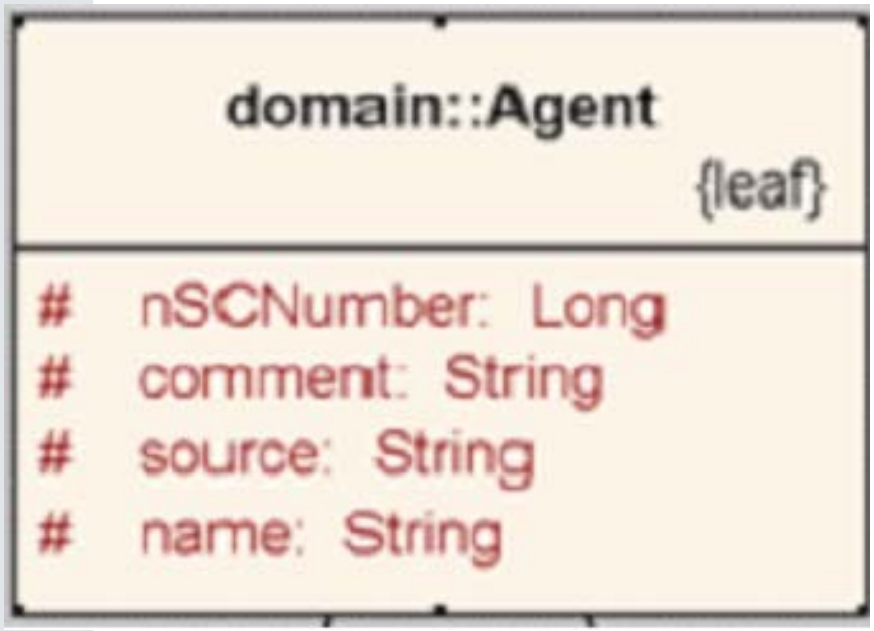
# Common Data Elements

Cancer Data Standard Repository (caDSR)


- 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
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
Disease_Has_Normal_Cell_Origin
Glandular Cell
Disease_Has_Normal_Tissue_Origin
Epithelium
Disease_Has_Primary_Anatomic_Site
Prostate Gland
Relationships
Information about this concept:
Preferred_Name
Prostate Adenocarcinoma
Semantic_Type
Neoplastic Process
Unified Medical Language System Concept Identifier
C0007112
Preferred Name
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
Definition
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
Links
NCI_The
Cancer Institute

# Tying it all together: The caCORE semantic management framework




| Metadata identifiers | Desc. Logic Concept Codes |
|---|---|
| 2223333v1 | C1708 |
| 2223866v1 | C1708:C41243 |
| 2223869v1 | C1708:C25393 |
| 2223870v1 | C1708:C25683 |
| 2223871v1 | C1708:C42614 |

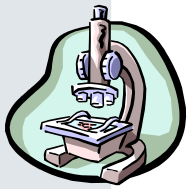
caDSR

**Bioinformatics Objects**

Common Data Elements

**Enterprise Vocabulary**

# caCORE Infrastructure wiring




Public APIs

Domain object metadata

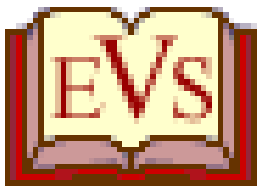
user application

caCORE SDK

Common data elements (CDEs)

Common data elements

Vocabulary for CDE specification



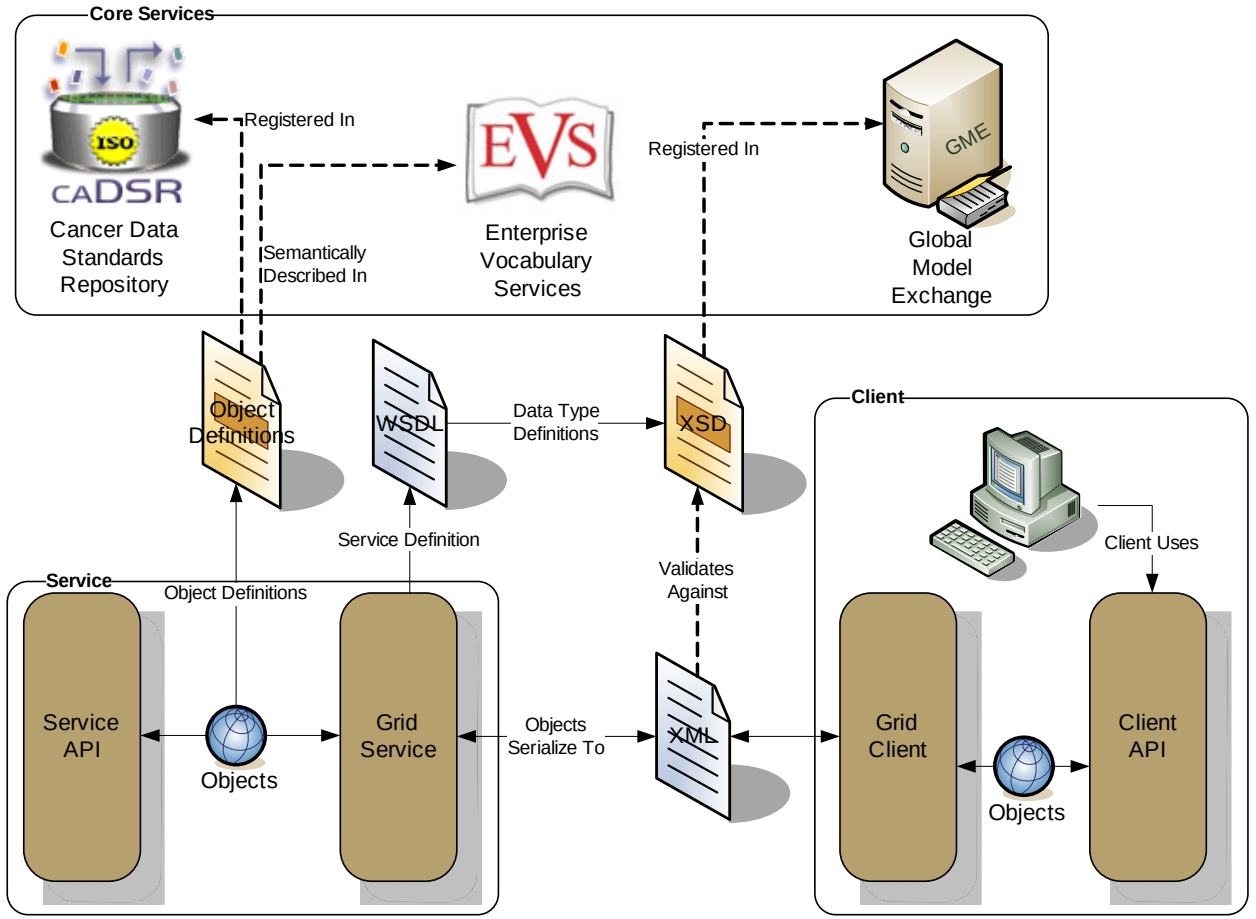

Dictionary, thesaurus services

# Cancer Bioinformatics Grid (caBIG) Use Cases

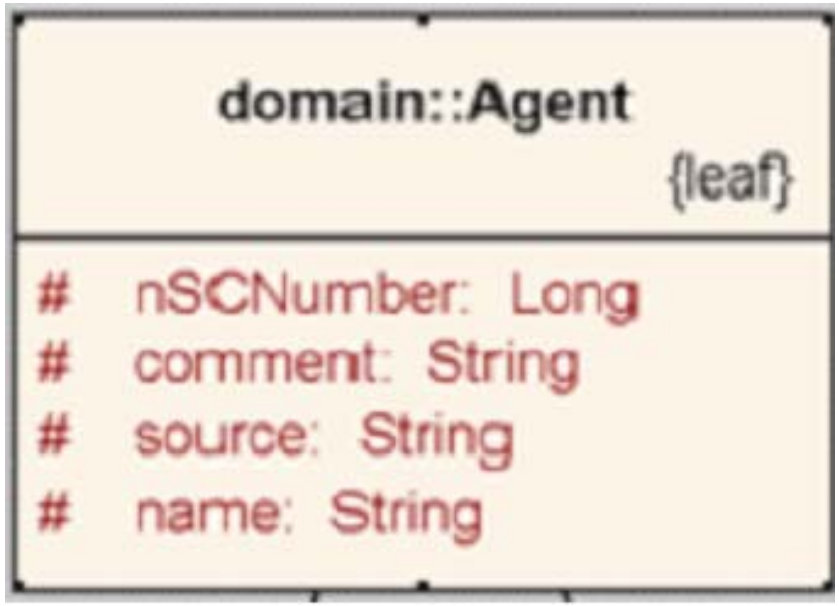
- **Advertisement**
  - **Service Provider** composes service metadata describing the data or analytic service and publishes it to grid.
- **Discovery**
  - **Researcher** (or application developer) specifies search criteria describing a service of interest
  - The research submits the discovery request 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


# 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)



# Semantic metadata example: Agent

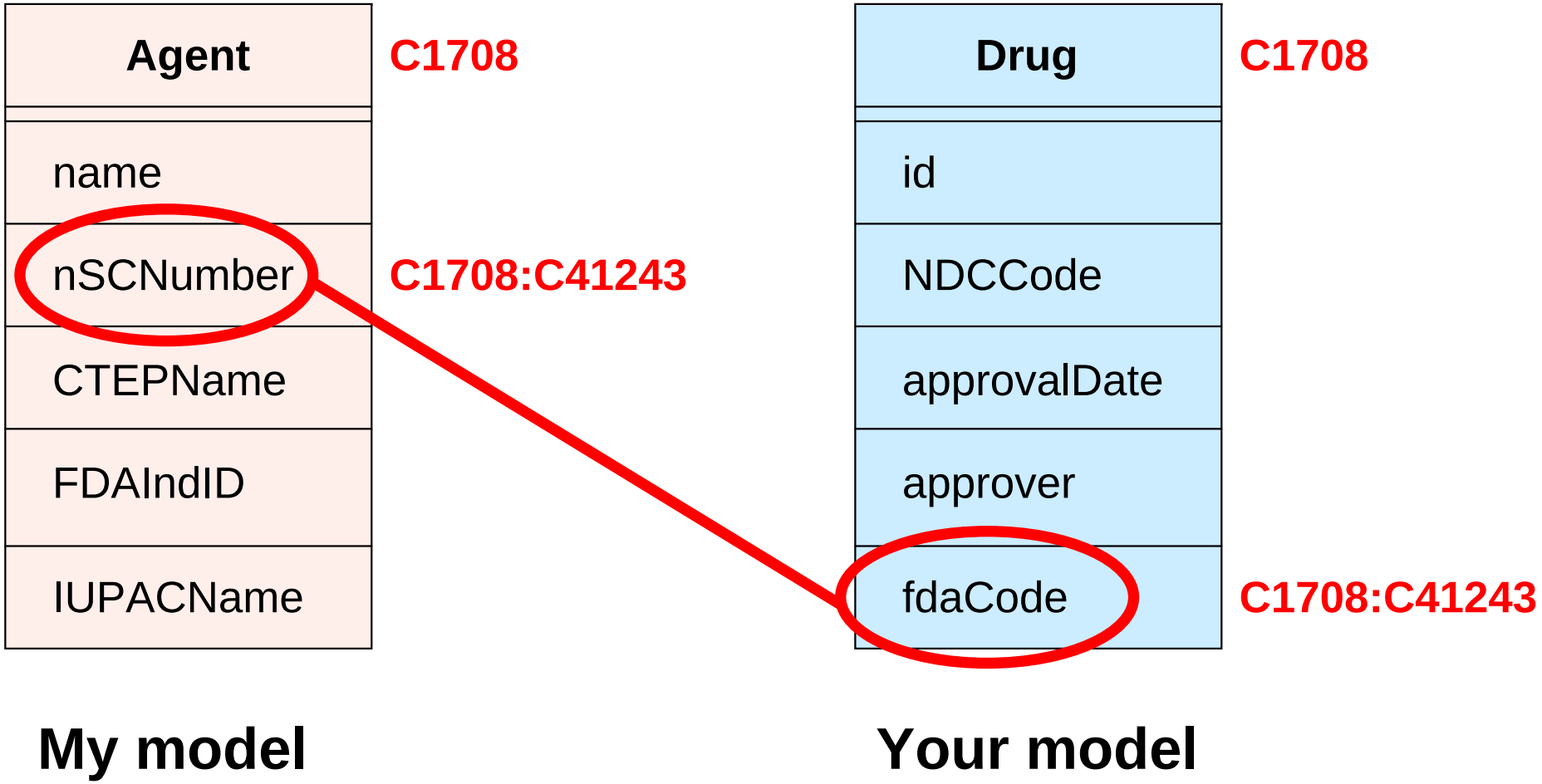



```
<Agent>
    <name>Taxol</name>
    <nSCNumber>007</nSC
      Number>
</Agent>
```

# 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 |

# Context Specific Computable Interoperability



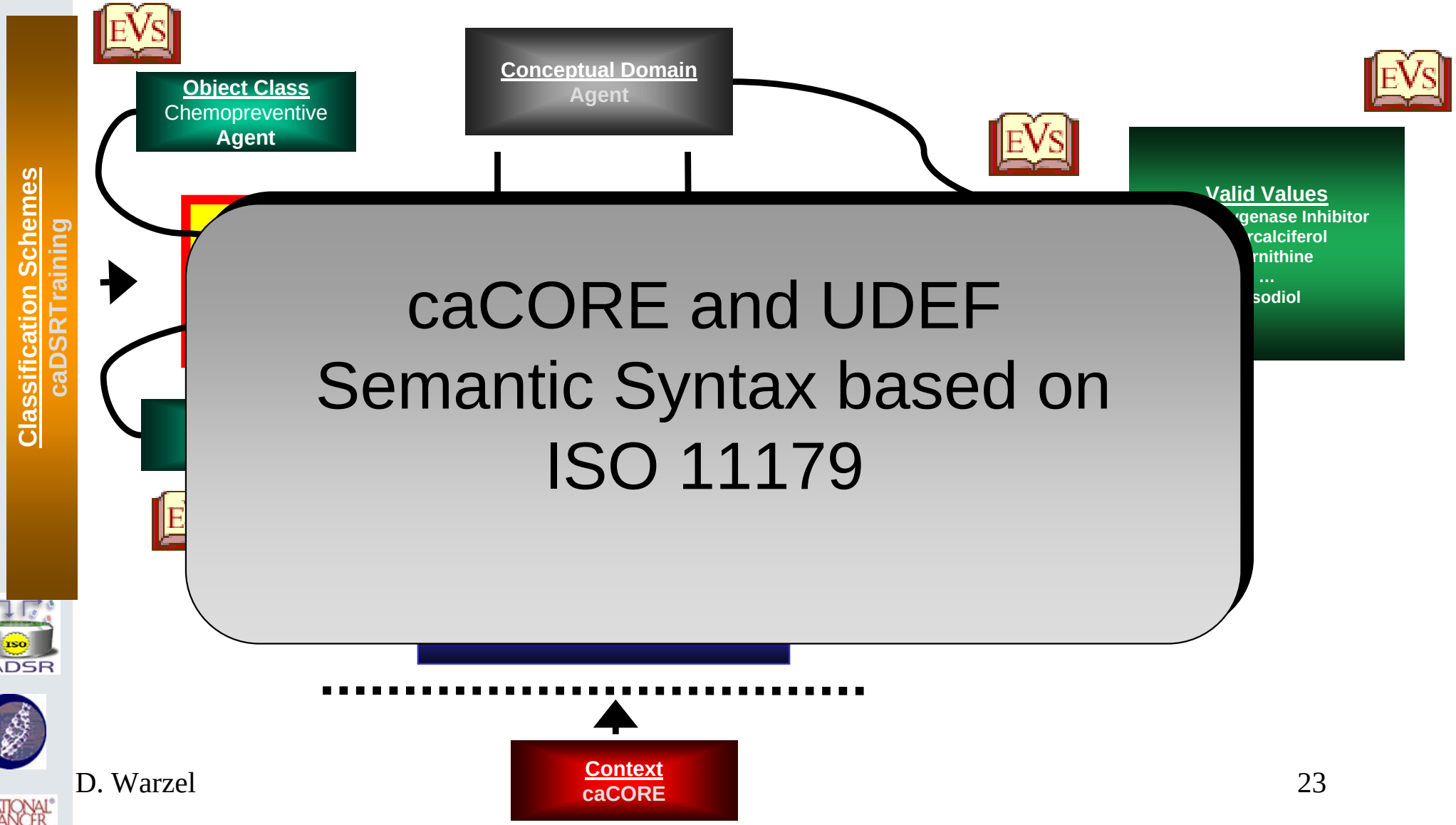
# Cancer Data Standards Registry (caDSR)

- 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 and Browser: automatically register UML models as metadata components, view, share, reuse
  - 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

# Convergence Scenario
# Why caCORE?

- Similar goals and objectives
  - Consolidated Health Informatics (CHI)
    - Register and utilize United States health data elements and vocabulary standards to create a semantic service oriented national health infrastructure
  - National Cancer Institute (NCI)
    - Register and utilize cancer data elements and vocabulary standards to create a semantic service oriented cancer research infrastructure


# caDSR Implementation of ISO/IEC 11179 MetaModel



# caDSR Metadata Registry

- Goals tools development:
  - Simplify development and creation of ISO/IEC 11179 compliant metadata by Data Element Curators and UML Modelers
  - Simplify consumption of Data Elements and standard vocabularies by end users and application developers through APIs and web services
  - Enhance reuse of Data Elements across domains
  - Enable semantic consistency across research domains
  - Support metadata life-cycle and governance processes
- Created, maintained by NCI Contractors and Open Development model
- Available as an open-source download

* Training *

# caCORE SDK Components


- **UML Modeling Tool** (any with XMI export)
- **Se** ut
- **U** ca
- **C** r)
- **S** Mo

caCORE SDK Generates semantically interoperable systems!


# 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

# New Partners

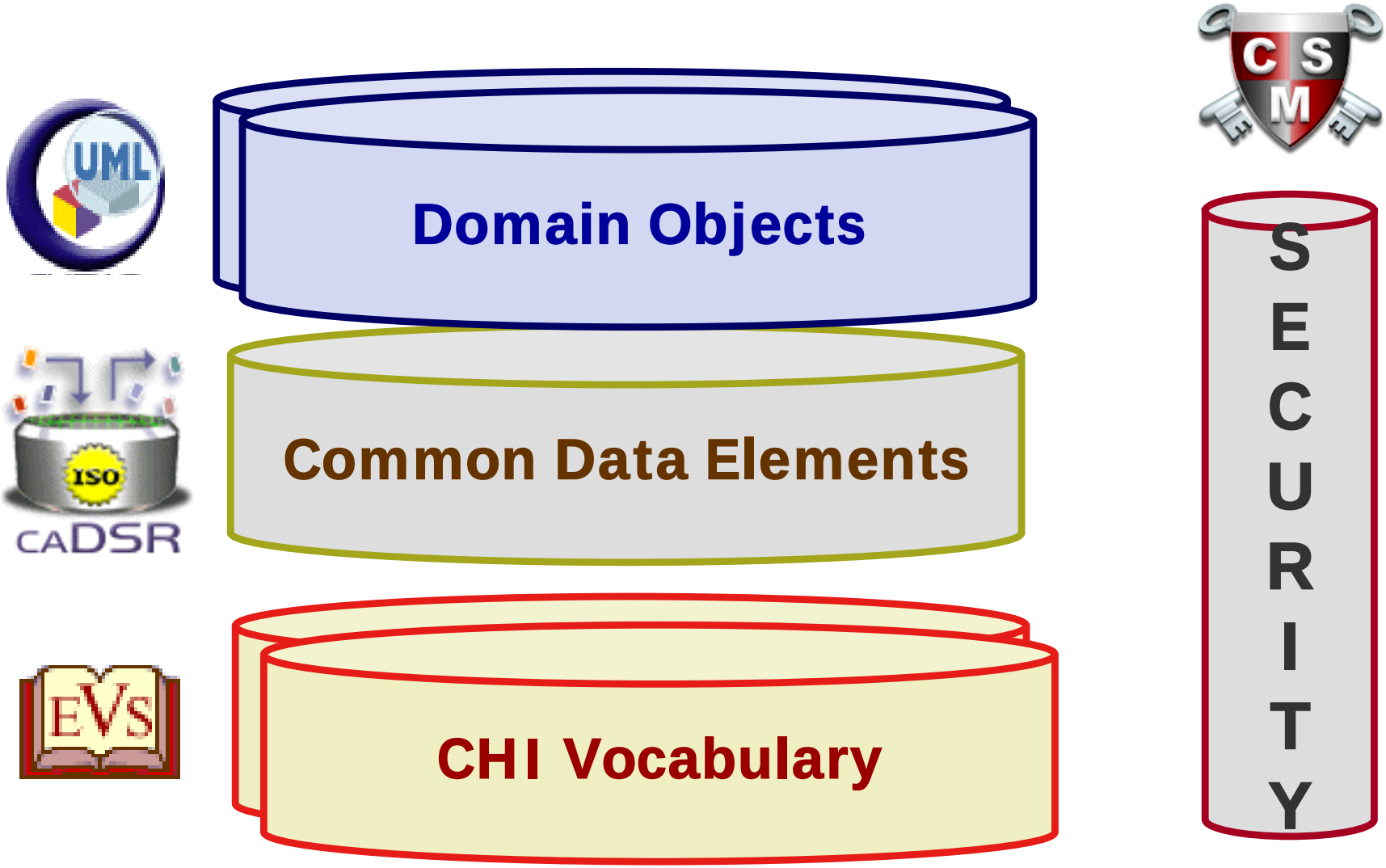
Planning/Implementation:

- National Icelandic Center for Oncology
  - Multi-lingual MDR
- HL7 Value Sets
- HL7 National Library of Medicine (NLM) Project
  - Register HL7 MDE mapped to HL7 vocabulary
- Department of Homeland Security

Exploring:

- National Institute of Neurological and Disorders and Syndromes (NINDS)
- National Cancer Research Institute UK (NCRI)

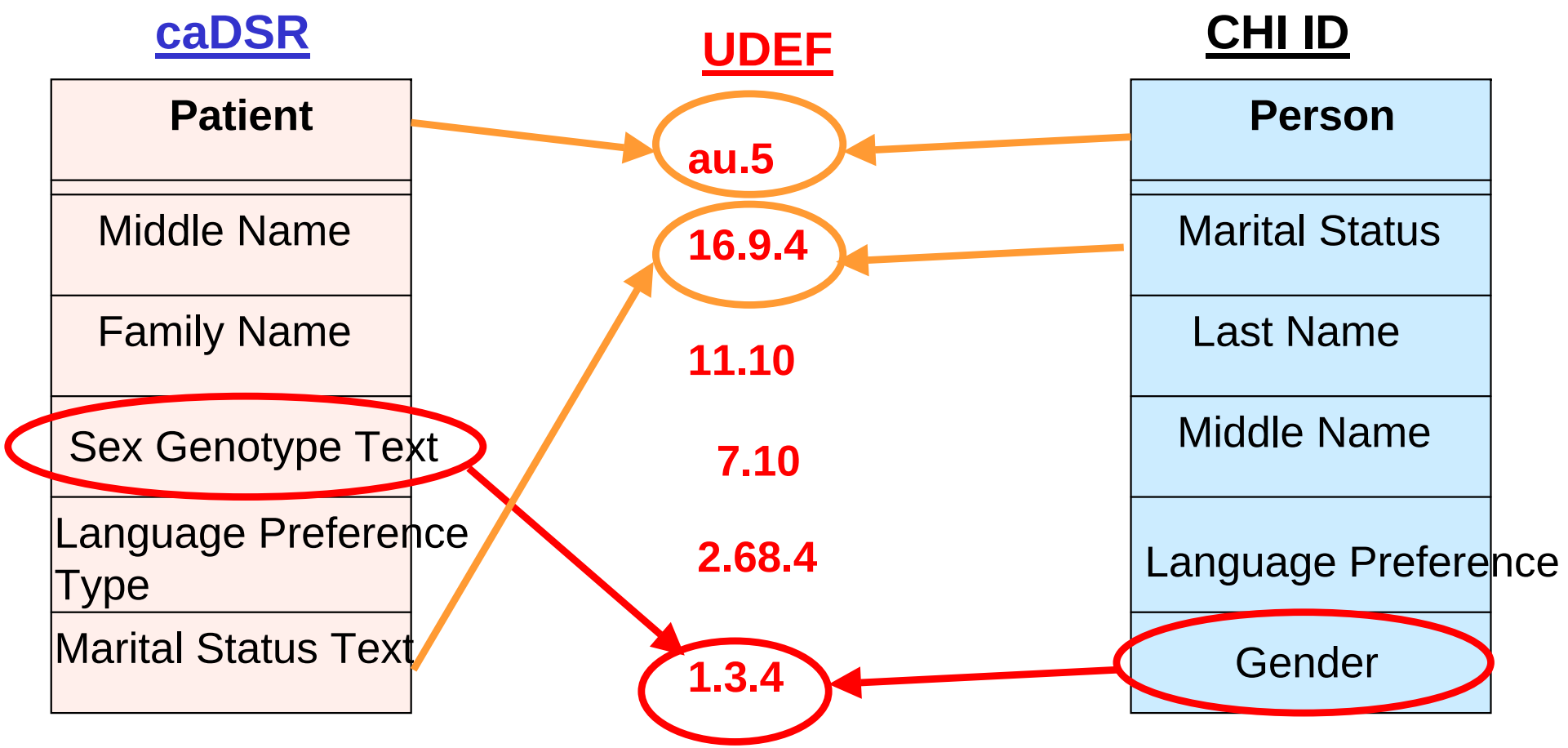
# Use Case chiCORE?



# Current Vocabularies

- NCI Thesaurus
  - HL7 registered Cancer specific
- NCI Metathesaurus
  - Based on NLM UMLS +
- LOINC
- SNOMED
- MeDRA
- VA NDF-RT
  - Veteran's Administration National Drug File Reference Terminology
- Gene Ontology (GO)


# UDEF
# Computable Interoperability?




**Patient Person Gender Genotype Code = au.5_1.3.4**

**Patient Person Marital Status Code = au.5_16.9.4**

# Documentation/Recommended Reading Materials

- caCORE Homepage:
  - http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview
- caCORE User Application Manual:
  - ftp://ftp1.nci.nih.gov/pub/cacore/NCICBapplications/NCICBAppManual.pdf
- caCORE Technical Guide:
  - ftp://ftp1.nci.nih.gov/pub/cacore/caCORE3.1_Tech_Guide.pdf – caCORE APIs
- caCORE Training
  - http://ncicb.nci.nih.gov/NCICB/training
- caDSR Business Rules
  - http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr/business_rules
- caDSR_Users List serv subscribe:
  - http://list.nih.gov
  - Send Request for caDSR Account to: ncicb@pop.nci.nih.gov
- caBIG home page: documentation about the Grid
  - http://cabig.nci.nih.gov

# 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

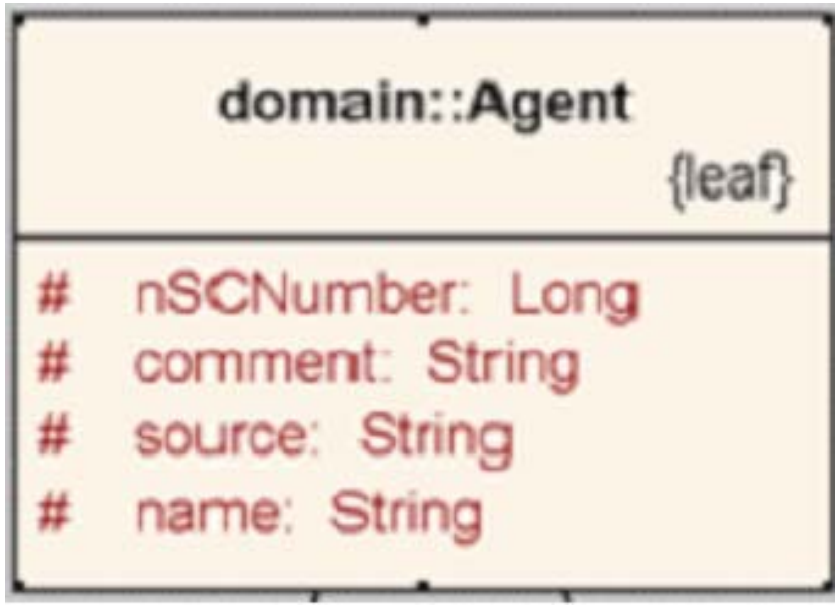
## Industry Partners

*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
- 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
  - Christophe Ludet
- BAH
  - Arumani Manisundaram

# Semantic metadata example: Agent




```
<Agent>
    <name>Taxol</name>
    <nSCNumber>007</nSC
      Number>
</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 |