National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Introduction to Interoperability and Compatibility

George A. Komatsoulis, Ph.D.

Director, Quality Assurance and Compliance

NCICB

National Cancer Institute

caBIG cancer Biomedical Informatics Grid

Courtesy: Charlie Mead

Interoperability

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

Syntactic interoperability

Semantic interoperability

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Some data on the Grid

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

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Attribute	Value	NCI Metadata	CIA Metadata
Agent		A chemical compound administered to a human being to treat an existing disease or condition, or prevent the onset of a disease or condition	A sworn intelligence agent; a spy
nSCNumber	007	Identifier given to a chemical compound by the US Food and Drug Administration (FDA) Nomenclature Standards Committee (NSC)	Identifier given to an intelligence agent by the National Security Council
Name	Taxol	Name of a chemical compound given by the NCI Cancer Therapeutics Evaluation Program (CTEP)	Code name given to intelligence agents by the Central Intelligence Agency (CIA)

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

caBIG™ Compatibility Guidelines

- The caBIG™ compatibility guidelines are designed to insure that systems designed in a Federated environment are still interoperable on the caBIG™ Grid, both syntactically and semantically
- Since achieving interoperability is a process, caBIG™ recognizes four levels of compatibility, starting from Legacy (not interoperable) through Bronze, Silver and Gold (fully interoperable)
- caBIG™ compatibility is all about interfaces rather than the scientific content of the system
- The analogy is to a city
  - In cities architects are free to design buildings that perform myriad functions and that take many distinct forms
  - Nevertheless, all buildings in the city are required to conform to certain specifications in order to receive electricity, water, steam, mail, etc.

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Maturity Model	Legacy	Bronze	Silver	Gold
Interface Integration	- No Programming interfaces to the system are available. Only local data files in a custom format can be read - Some ad hoc data transfer mechanism such as FTP	- Provide baseline* programmatic access to data. Data can be read from remote electronic sources or from commonly used file formats Data can be pushed out to from applications to other external data sources	- Well-described API's that provide access to data objects. - System architecture separated into tiers and interoperable components - Data read in from standards-based electronic sources that support standard or commonly used interchange formats - Documented component description of the underlying data structures that are accessible - Standard messaging systems where appropriate	- All features of Silver, plus: - Interoperable with data grid architecture to be defined by caBIG - Fully componentized provide access to individual resources in the form of grid services
Vocabularies / Terminologies & Ontologies	- Free text used throughout for data collection	- Use of publicly accessible standardized controlled vocabularies as well as local terminologies	- Standard terminologies approved by public standards bodies or the caBIG Vocabulary/CDE Workspace are used for all relevant data collection fields.	- All features of Silver, plus: - Fully compliant with caBIG recommended standards for vocabulary terminology services and content sources
Data Elements	- No Structured metadata is recorded	- Some type of metadata describing the information in the system is used for data collection and external reporting. Metadata is retrieved from external repository shared by multiple applications. - Common Data Elements should be built using controlled terminology	- Use common standard electronic representation for CDE's such as ISO 11179 or comparable standard - CDEs are harmonized and re-used from across the Domain Workspace - Common Data Elements are built using standard controlled terminologies approved by public standards bodies or the caBIG Vocabulary/CDE Workspace	- All features of Silver, plus: - Programmatic access to all metadata, including data class descriptions, site and source information, and any other caBIG-defined metadata requirements and use information models - Use the caBIG standard or electronic representation of metadata and Common Data Elements
Information Models	- No particular information model is used to represent data	- Some type of diagrammatic model describing the data relationship is available in electronic format	- Information models defined in a standard modeling language such as UML	- All features of Silver, plus: - Information models are harmonized with other s across the caBIG Domain Workspace

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

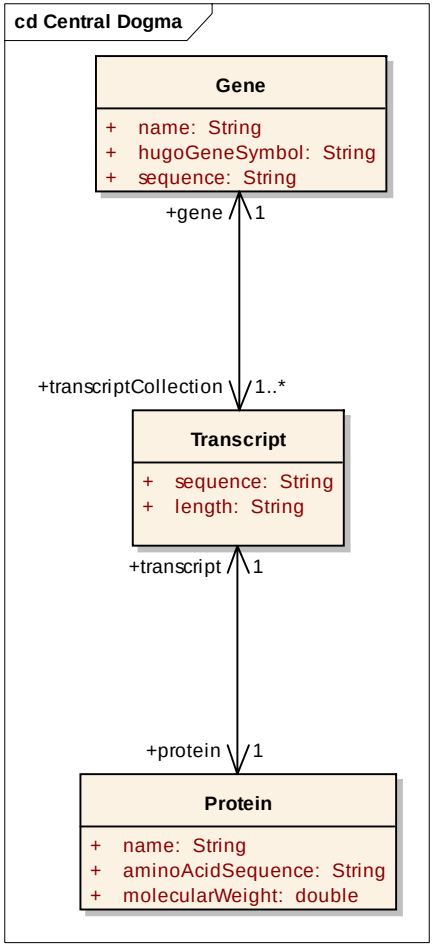
Syntactic Interoperability

- The solution for syntactic interoperability in caBIG is for all systems to provide an Object Oriented Application Programmer Interface.
- Object Oriented Interfaces can be implemented in many programming languages.
- This interface can be connected to the caGrid so that the local data repository is globally accessible in a language independent way.
- The interface is described by an information model, which acts as the junction between the syntactic components and the semantic components.

National Cancer Institute

caBIG cancer Biomedical Informatics Grid

Domain Information Modeling



- An Domain Information Model is a representation of our understanding of an area of knowledge.
- Domain Information Models consist of ‘Classes’ that represent ‘things’ in the real world
- Classes contain ‘attributes’ that are characteristics of different instances of things in the real world.
- Relationships between the classes are described by ‘associations’ and indicated by lines with directionality and cardinality
- Each class plus attribute creates one Common Data Element (CDE)

National Cancer Institute

caBIG cancer Biomedical Informatics Grid

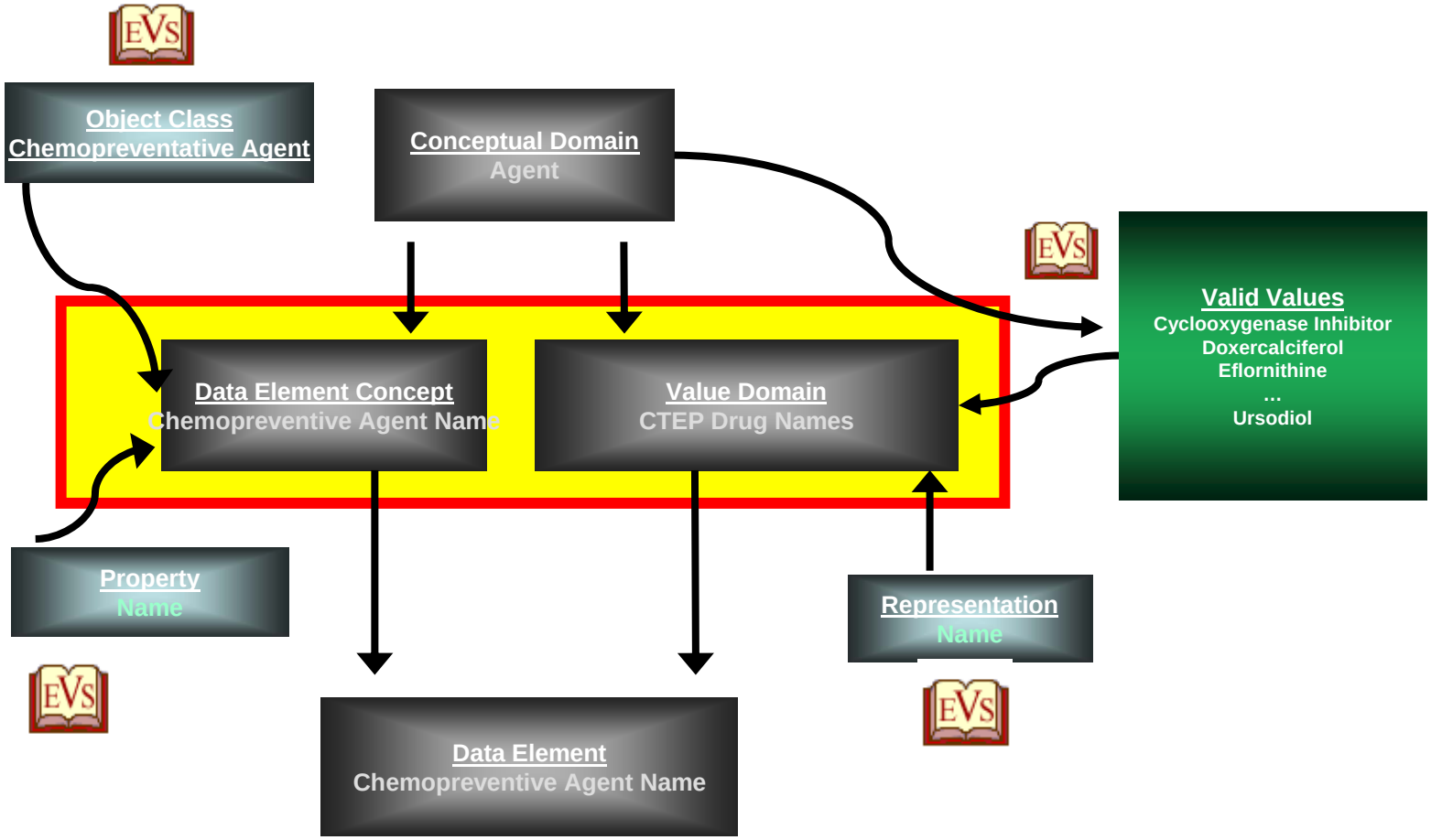
Cancer Data Standards Repository (caDSR)

- Basic caDSR unit of metadata information to describe a datum is a Common Data Element or CDE
- Enterprise-class system for storing metadata, with APIs that give runtime access to both metadata and semantics
- Implements the ISO 11179 standard, a flexible model for describing arbitrary metadata
- Used to describe metadata associated with clinical case report forms and UML Models

National Cancer Institute

caBIG cancer Biomedical Informatics Grid

caDSR Implementation of ISO/IEC 11179 Model



National Cancer Institute

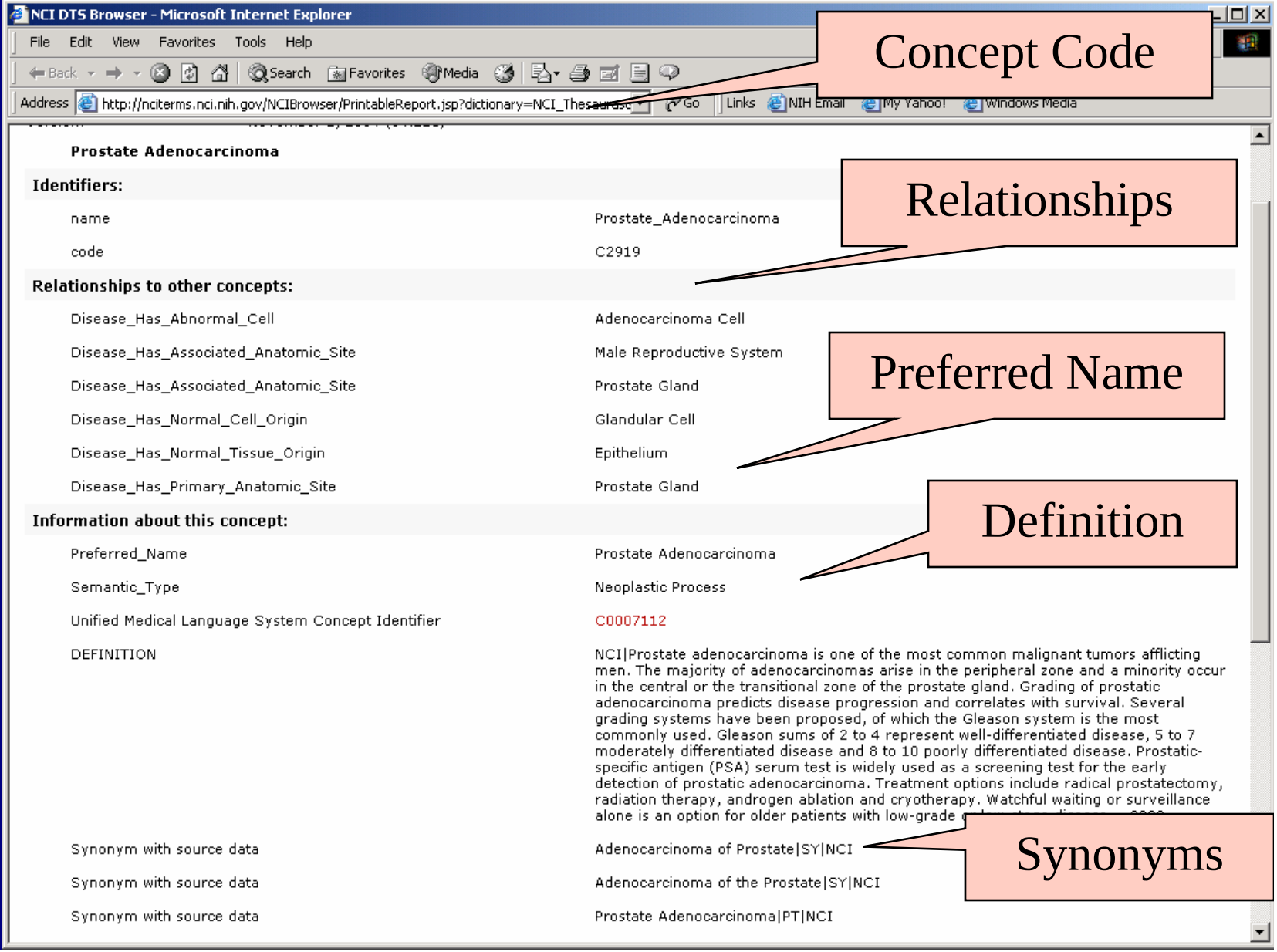
caBIG™ cancer Biomedical Informatics Grid™

Enterprise Vocabulary Services

- Controlled vocabulary resources for caCORE and the cancer research community
- Vocabulary Products and Services
  - NCI Thesaurus
  - NCI Metathesaurus
  - External vocabularies
- NCI Thesaurus - controlled vocabulary source for metadata
  - Has excellent coverage of cancer terminology
  - Expands based on needs for additional terminology
  - Based on concepts rather than terms
  - Each concept has a unique identifier or CUI with definitions and synonym

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

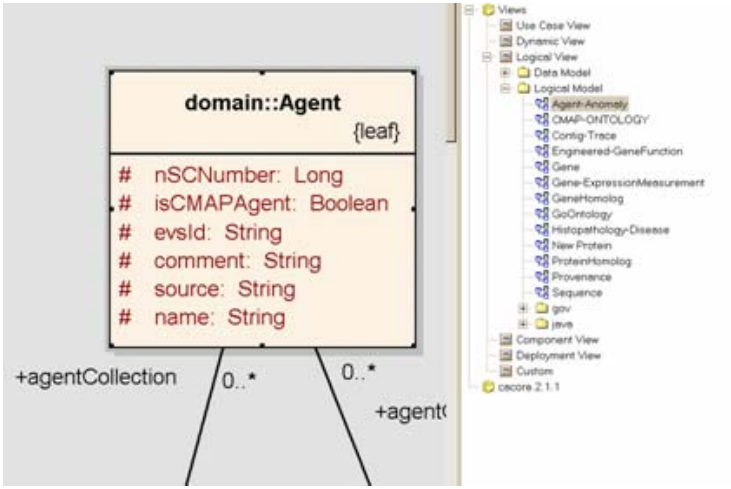


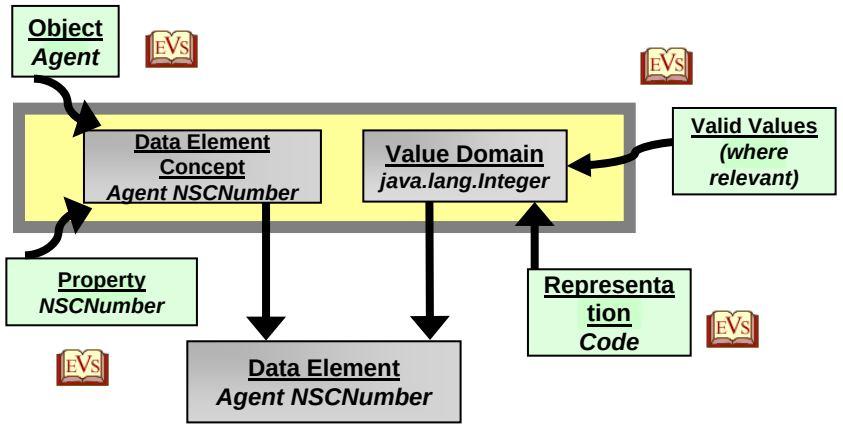
National Cancer Institute

caBIG cancer Biomedical Informatics Grid

Semantic Integration in caBIG

- **UML Class = ISO Object Class**
  - Example: Agent = C1708 (Agent)
- **UML Class Attribute = ISO Property**
  - Example: nSCNumber = C41243 (NSC Code)
- **UML Class + UML Class Attribute = ISO Data Element Concept**
  - Agent:nSCNumber = C1708:C41243
- **UML (Java) Datatype = ISO Value Domain (at least initially)**
  - Example: java.lang.Integer
- **UML Class + UML Class Attribute + Datatype/Valid Values = ISO Common Data Element**
  - Example: Agent:nSCNumber: java.lang.Integer

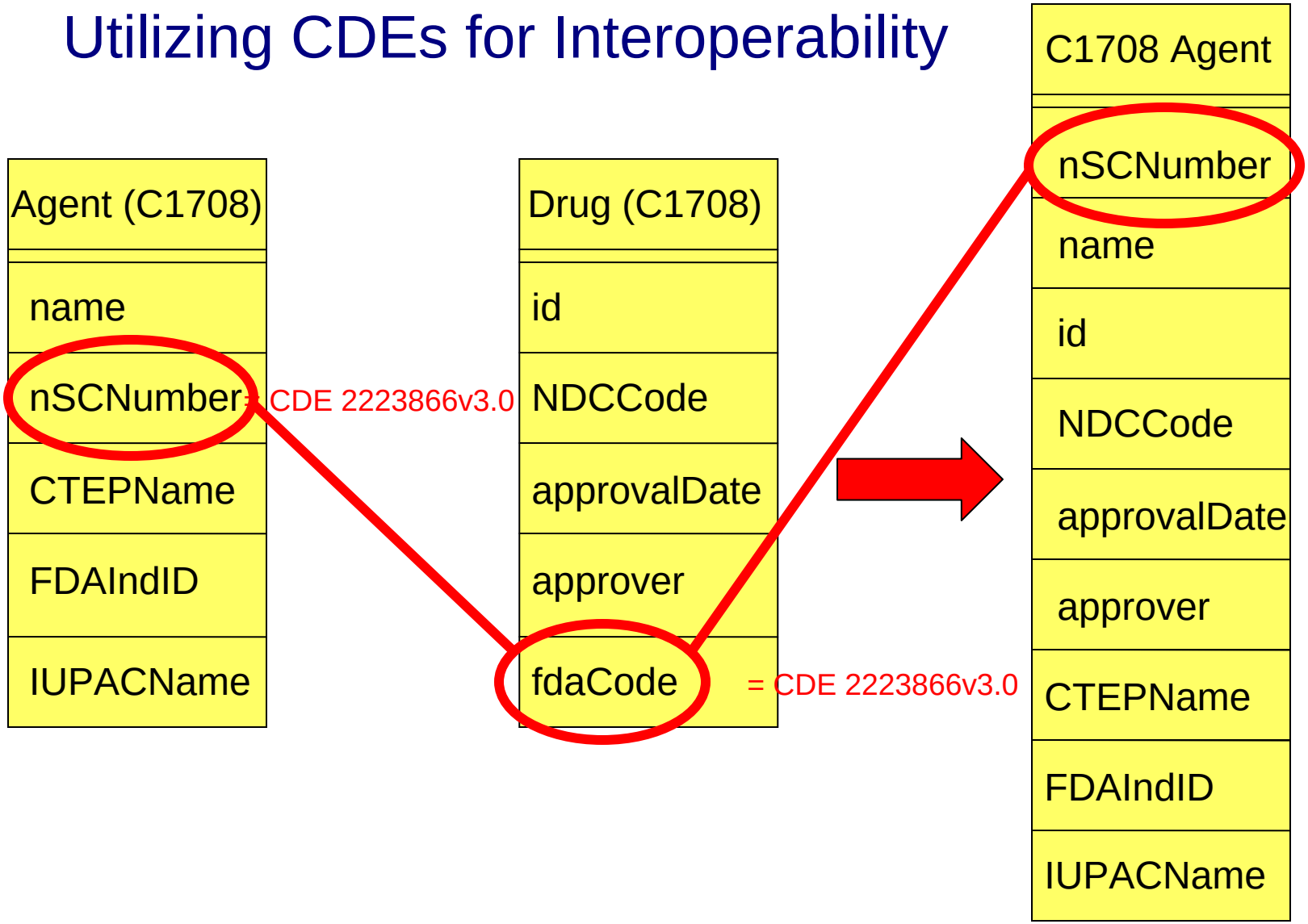




National Cancer Institute

caBIG cancer Biomedical Informatics Grid

Utilizing CDEs for Interoperability



National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Maturity Model	Legacy	Bronze	Silver	Gold
Interface Integration	- No Programming interfaces to the system are available. Only local data files in a custom format can be read - Some ad hoc data transfer mechanism such as FTP	- Provide baseline* programmatic access to data. Data can be read from remote electronic sources or from commonly used file formats Data can be pushed out to from applications to other external data sources	- Well-described API's that provide access to data objects. - System architecture separated into tiers and interoperable components - Data read in from standards-based electronic sources that support standard or commonly used interchange formats - Documented component description of the underlying data structures that are accessible - Standard messaging systems where appropriate	- All features of Silver, plus: - Interoperable with data grid architecture to be defined by caBIG - Fully componentized provide access to individual resources in the form of grid services
Vocabularies / Terminologies & Ontologies	- Free text used throughout for data collection	- Use of publicly accessible standardized controlled vocabularies as well as local terminologies	- Standard terminologies approved by public standards bodies or the caBIG Vocabulary/CDE Workspace are used for all relevant data collection fields.	- All features of Silver, plus: - Fully compliant with caBIG recommended standards for vocabulary terminology services and content sources
Data Elements	- No Structured metadata is recorded	- Some type of metadata describing the information in the system is used for data collection and external reporting. Metadata is retrieved from external repository shared by multiple applications. - Common Data Elements should be built using controlled terminology	- Use common standard electronic representation for CDE's such as ISO 11179 or comparable standard - CDEs are harmonized and re-used from across the Domain Workspace - Common Data Elements are built using standard controlled terminologies approved by public standards bodies or the caBIG Vocabulary/CDE Workspace	- All features of Silver, plus: - Programmatic access to all metadata, including data class descriptions, site and source information, and any other caBIG-defined metadata requirements and use information models - Use the caBIG standard or electronic representation of metadata and Common Data Elements
Information Models	- No particular information model is used to represent data	- Some type of diagrammatic model describing the data relationship is available in electronic format	- Information models defined in a standard modeling language such as UML	- All features of Silver, plus: - Information models are harmonized with other s across the caBIG Domain Workspace

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Architecture Workspace

- The Architecture workspace is charged with building the caBIG™ Grid and ensuring syntactic interoperability
  - Develop caBIG™ compatibility guidelines in the area of interface integration.
  - Provide mentors to caBIG™ funded development projects in the area of interface integration
  - Create caGrid, the caBIG™ Grid
  - Perform interoperability reviews of caBIG™ funded development projects to ensure compliance with caBIG™ compatibility guidelines for interface integration

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Vocabulary and Common Data Elements Workspace

- The VCDE workspace is charged with ensuring semantic interoperability
  - Develop caBIG™ compatibility guidelines in the areas of information modeling, metadata and vocabularies
  - Provide mentors to caBIG™ funded development projects in the area of information modeling, metadata and vocabularies
  - Facilitate development of caBIG™ data standards, both vocabularies and CDEs
  - Perform interoperability reviews of caBIG™ funded development projects to ensure compliance with caBIG™ compatibility guidelines

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

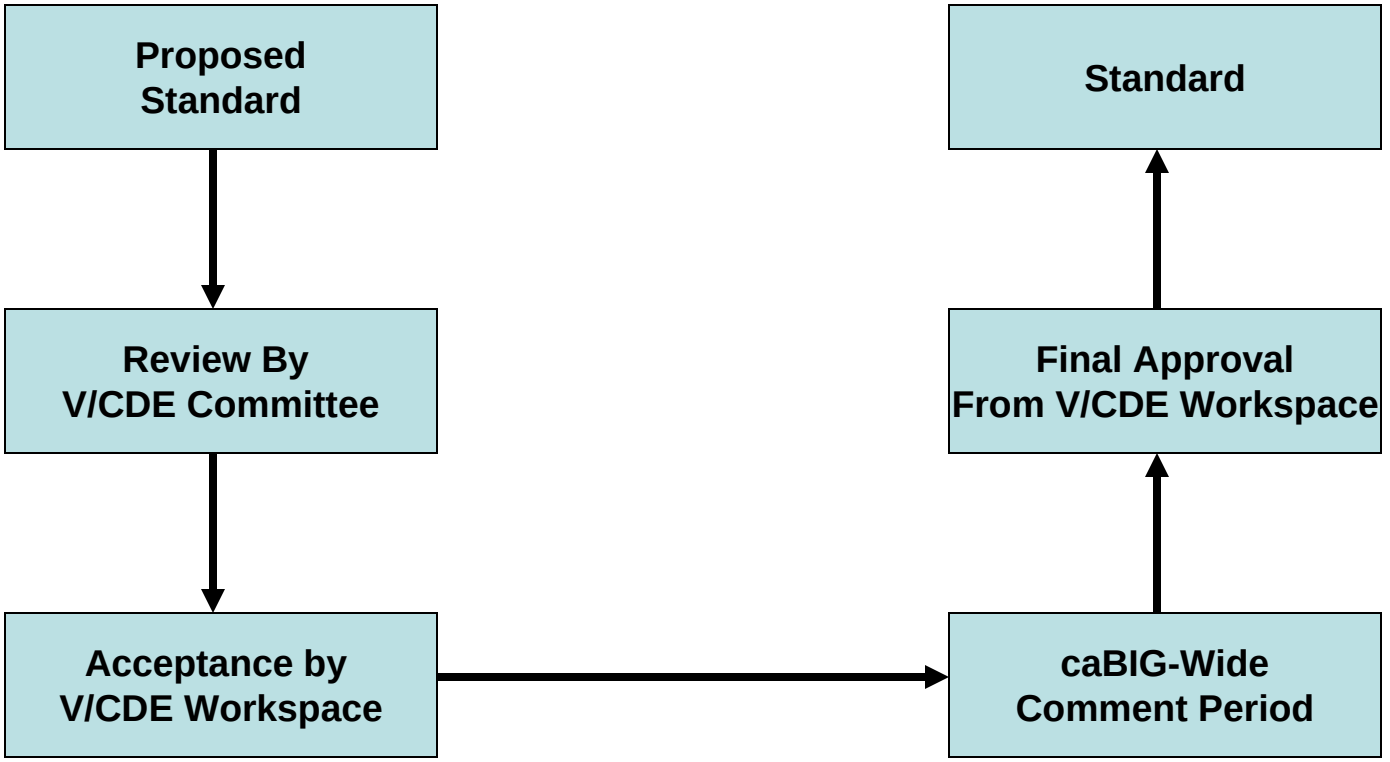
Data Standards in caBIG™

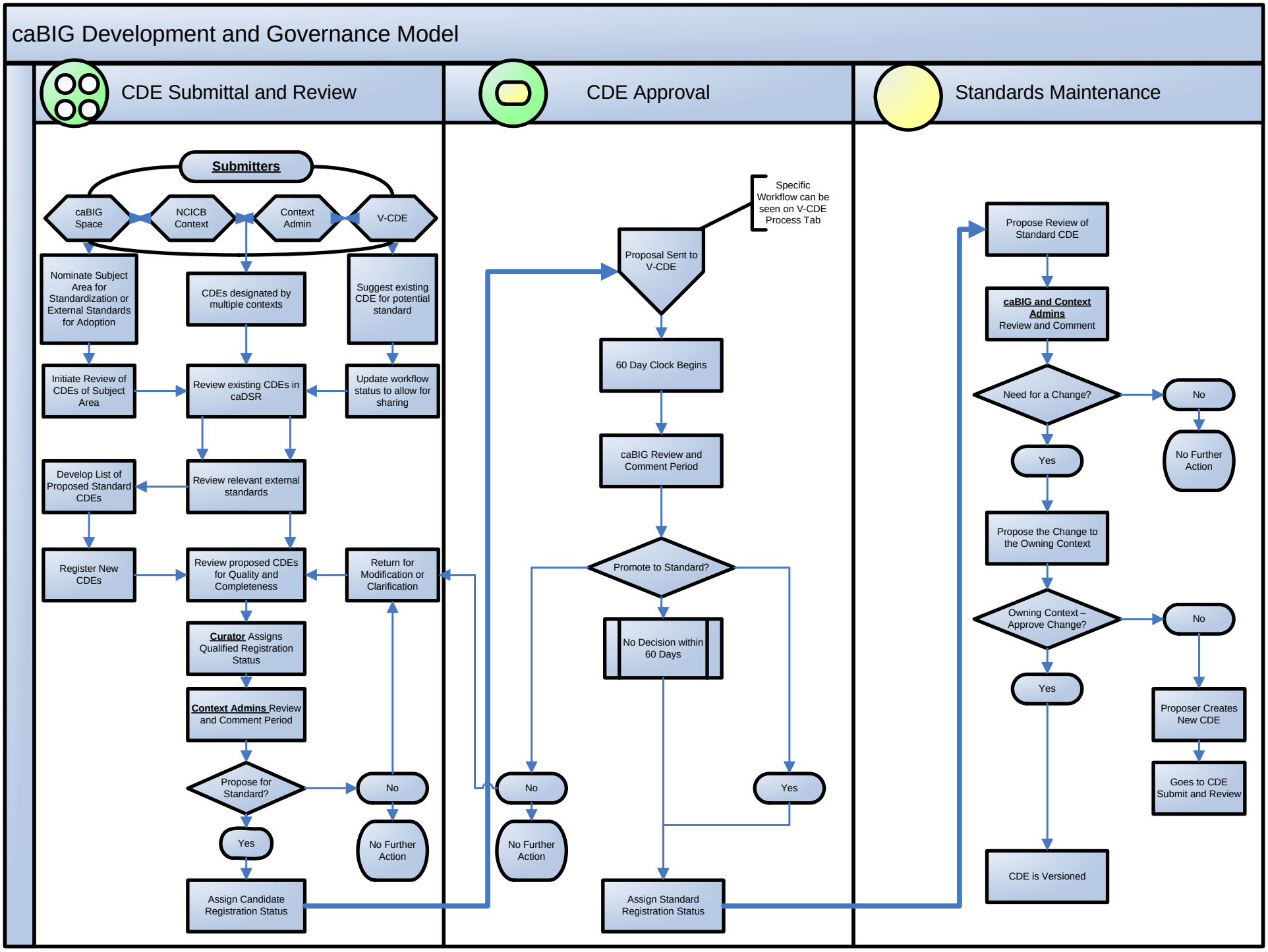
- The V/CDE workspace is responsible for facilitating the development and ratification of Data Standards for caBIG™
- Data Standards can be Vocabularies or Common Data Elements (CDEs) with their associated controlled terminology
- A caBIG™ Data Standard is, in effect, a 'pre-approved' mechanism for semantically modeling an attribute or series of attributes in a data object. Ideally, having a standard available shortens development time for other projects that need to present such data
- Whenever possible, caBIG™ adopts standards that are derived from other standards bodies (HL7, ISO, USPS, UPU, W3C, etc.) and in general use within our community
- In the last year, the V/CDE workspace has developed a consensus driven mechanism for approving Data Standards and applied it to an increasing number of CDEs

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Data Standards Process




caBIG Development and Governance Model
CDE Submittal and Review
CDE Approval
Standards Maintenance
Submitters
caBIG Space
NCICB Context
Context Admin
V-CDE
Nominate Subject Area for Standardization or External Standards for Adoption
CDEs designated by multiple contexts
Suggest existing CDE for potential standard
Initiate Review of CDEs of Subject Area
Review existing CDEs in caDSR
Update workflow status to allow for sharing
Develop List of Proposed Standard CDEs
Review relevant external standards
Register New CDEs
Review proposed CDEs for Quality and Completeness
Return for Modification or Clarification
Curator Assigns Qualified Registration Status
Context Admins Review and Comment Period
Propose for Standard?
No
No Further Action
Yes
Assign Candidate Registration Status
No
No Further Action
Proposal Sent to V-CDE
Specific Workflow can be seen on V-CDE Process Tab
60 Day Clock Begins
caBIG Review and Comment Period
Promote to Standard?
No Decision within 60 Days
Yes
Assign Standard Registration Status
Propose Review of Standard CDE
caBIG and Context Admins Review and Comment
Need for a Change?
No
No Further Action
Yes
Propose the Change to the Owning Context
Owning Context – Approve Change?
No
Proposer Creates New CDE
Goes to CDE Submit and Review
Yes
CDE is Versioned

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

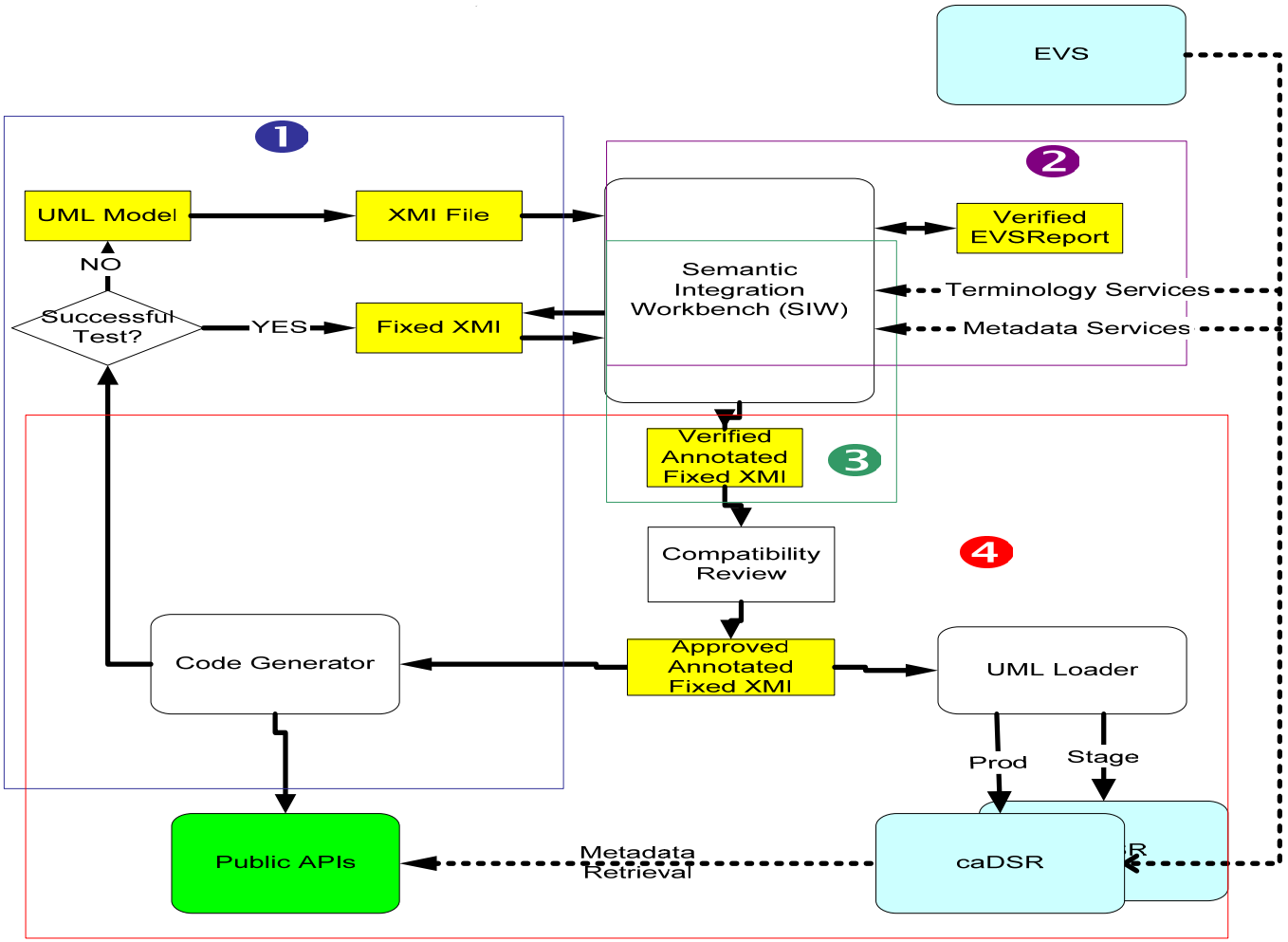
Enabling Technology

- The NCI provides freely available enabling technology for caBIG™ compatibility
- These technologies are distributed under a ‘non-viral’ open source license.
- caCORE
  - Enterprise Vocabulary Services (EVS)
  - Cancer Data Standards Repository (caDSR)
- caCORE Software Development Kit
  - When complete process is followed, the outcome is a caBIG ‘Silver’ compliant data system.

National Cancer Institute

caBIG cancer Biomedical Informatics Grid

Semantic Integration Process



National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

Grid Technology in caBIG™

- What is a 'Grid'
  - "A Grid is a system that coordinates resources that are not subject to centralized control using standard, open, general-purpose protocols and interfaces to deliver nontrivial qualities of service." - Ian Foster **Grid Today**, July 20, 2002
- Grid Technology supplies two useful components to a network of computers:
  - Advertising: Inform the network about the capabilities of new systems
  - Discovery: Allow users to find resources that meet their needs.
- The caGrid project is the 'Grid in caBIG™'; the actual infrastructure that data and analytical services will use to interoperate.
- The current caGrid is version 0.5; construction of caGrid 1.0 is underway.
- The combination of data and analytical service nodes in caBIG™ produced a design that utilizes a variety of standard Grid technologies including the Globus Toolkit and OGSA-DAI, DQP, GRAM, etc.

National Cancer Institute

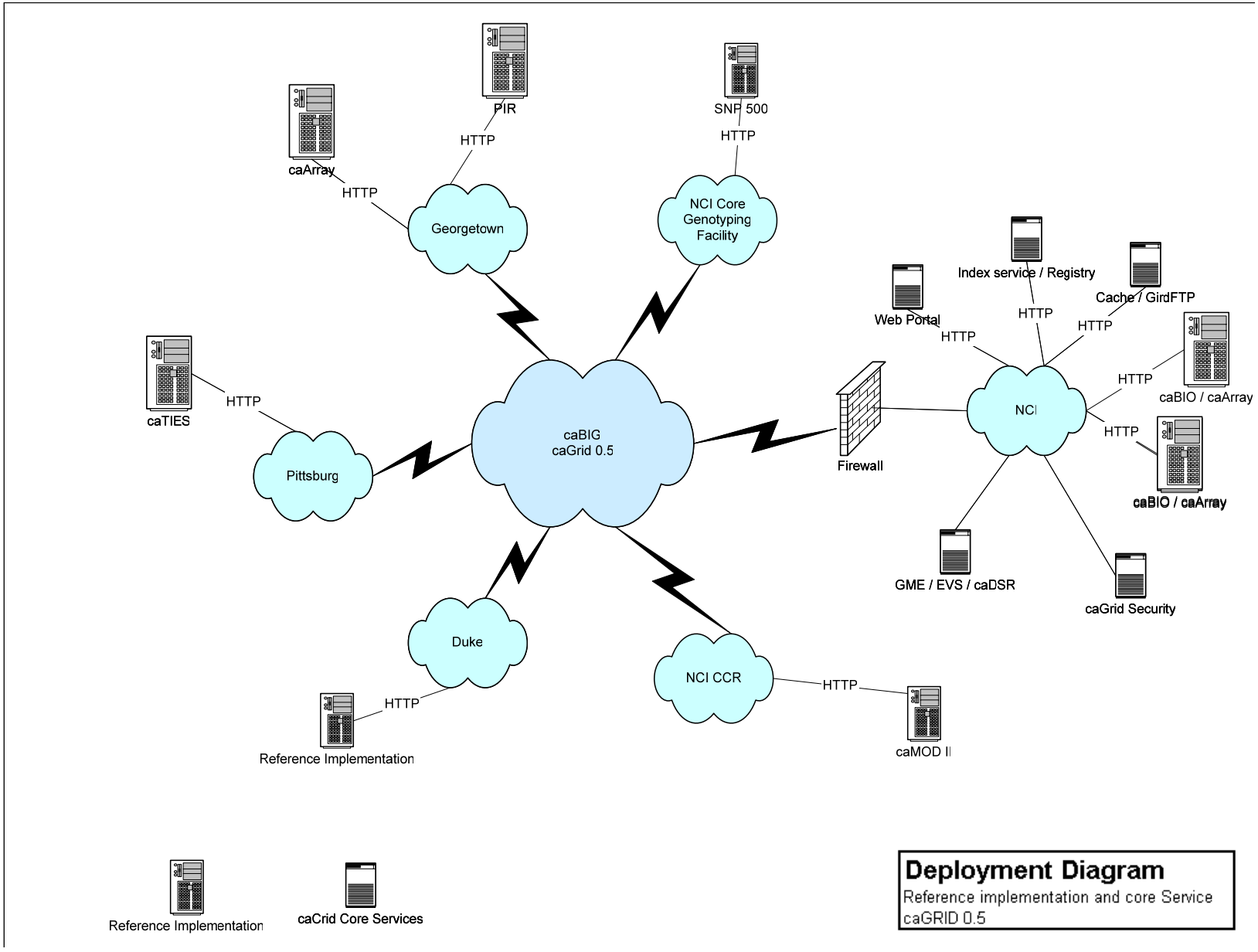
caBIG™ cancer Biomedical Informatics Grid™

caGrid 0.5 Services

- **Data Services**
  - **caBIO:** Gene-centric bioinformatics objects
    - NCICB-Rockville, MD
  - **caArray:** MAGE-OM compliant microarray repository
    - NCICB-Rockville, MD
    - Lombardi Cancer Center-Georgetown, DC
  - **gridPIR:** Protein Information Resource
    - Lombardi Cancer Center-Georgetown, DC
  - **caTIES:** Text Information Extraction System for pathology reports
    - UPMC-Pittsburgh, PA
  - **SNP500:** Polymorphism database with population frequencies
    - NCI Core Genotyping Facility-Gaithersburg, MD
  - **caMOD II:** Cancer Model Organism Database
    - NCI Mouse Models of Human Cancer Consortium (MMHCC)
- **Analytical Service**
  - **RProteomics:** Statistical analysis of proteomics data
    - Duke-Durham, NC

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™



National Cancer Institute

caBIG cancer Biomedical Informatics Grid

Certification of Compliance

- Projects that are funded by caBIG are reviewed for compliance with compatibility guidelines by the cross-cutting architecture and VCDE workspaces.
- Projects that are not funded by caBIG can participate in the new Bronze Compatibility Certification Program. The Bronze Certification Program is a verified self test.
- There will be a session on the Bronze Certification Program on Tuesday at 10:15AM
- Information about the Bronze program is available at https://cabig.nci.nih.gov/guidelines_compatibility/bronze/

National Cancer Institute

Acknowledgments

- **caBIG™** Program
  - Ken Buetow
  - Peter Covitz
  - Denise Warzel
  - Leslie Derr
  - Mary Jo Deering
  - Dianne Reeves
  - Jill Hadfield
  - Mark Adams
- V/CDE Leadership
  - Brian Davis
  - Michael Keller
- Architecture Leadership
  - Avinash Shanbhag
  - Manisundarum Arumani
- NCI Enterprise Vocabulary Services
  - Frank Hartel
  - Sherry De Coronado
  - Gilberto Fragoso
  - Larry Wright
  - Margaret Haber

National Cancer Institute

caBIG™ cancer Biomedical Informatics Grid™

- NCI Context Administrators
  - Ann Setser (CTEP)
  - Brian Campbell (CTEP)
  - Bev Meadows (DCP)
  - Chitra Mohla (DCEG/DCCPS)
- Ohio State University
  - Scott Oster
  - Shannon Hastings
  - Steve Langella
  - Tahsin Kurc
  - Joel Saltz
- Oracle
  - Steve Alred
  - Christophe Ludet
- TerpSys
  - Gavin Brennan
  - Troy Smith
  - Wei Lui
  - Doug Kanoza
- SAIC
  - Kathleen Gundry
  - Tommie Curtis
  - Brenda Maeske
  - Mary Cooper
  - Nafis Zebarjadi
  - William Sanchez
  - Tara Akhavan
  - Manav Kher
  - Rouwei Wu
  - Jijin Yan
- Panther Informatics
  - Brian Gilman
  - Nick Encina
- caBIG™ Patient Advocates
- caBIG™ Participants