caBIG cancer Biomedical Informatics Grid

Silver and Gold: Architecture, Compatibility and the Cancer Biomedical Informatics Grid

Peter A. Covitz, Ph.D.
Director, Bioinformatics Core Infrastructure
NCI Center for Bioinformatics

NATIONAL CANCER INSTITUTE

NCI Center for Bioinformatics

- Mission

  Provide infrastructure and applications to support and enhance the value of cancer research consortia

Partnerships Drive the NCICB

- Our friends in need…
  - – Program managers
  - – Consortium directors
  - – Genomics initiatives
  - – Clinical trial study groups
  - – NCI divisions
  - – Cancer centers [caBIG]
    - Major program to connect NCI-designated Cancer Centers across the nation to a common cancer biomedical informatics “grid”

Some Current Partners

- Genomics
  - Cancer Genome Anatomy Project (CGAP)
  - NCI Laboratory of Population Genetics
  - NCI microarray consortia, MGED
- Clinical Trials and Epidemiology Studies
  - NCI Center for Cancer Research, Division of Cancer Prevention, Cancer Therapy Evaluation Program
  - NCI Division of Cancer Control and Population Sciences, Division of Cancer Epidemiology and Genetics
  - SPOREs, Rembrandt, and other translational research trials
- Model Systems and Imaging
  - Mouse Models of Human Cancer Consortium
- Vocabulary and Data Standards
  - NCI Office of Communication
  - Several NCI Divisions
  - NLM, FDA, VA, other federal agencies

caBIG: From Village to City

W. 29th St
W. 28th St
W. 27th St
CHELSEA PARK
W. 26th St
W. 25th St
W. 24th St
W. 23rd St
W. 22nd St
CHELSEA
W. 21st St
W. 20th St
W. 19th St
W. 18th St
W. 17th St
W. 16th St
W. 15th St
W. 14th St
W. 13th St
Little W. 12th St
W. 12th St
W. 11th St
W. 10th St
W. 9th St
W. 8th St
12th Av
11th Av
10th Av
9th Av
8th Av
7th Av
5th Av
Madison Av
Broadway
Av of the Americas (6th Av)
MADISON SQUARE PARK
THEODORE ROOSEVELT BIRTHPLACE
West Side Expwy
Gansevoort St
Horatio St
Jane St
Hudson St
Washington St
Bethune St
Bank St
W. 11th St
Perry St
Charles St
W. 10th St
Christopher St
Greenwich St
Barrow St
Morton St
Leroy St
Clarkson St
W. Houston St
King St
Charlton St
WEST VILLAGE
Waverly Pl
Greenwich Av
Patchin Pl
Gay St
4th St
Bleecker St
Grove St
Bedford St
Commerce St
Jones St
Cornelia St
Carmine St
Downing St
WALKER PARK
Washington Sq N.
Washington Pl
WASHINGTON SQUARE
Washington Sq S.
W. 3rd St
MacDougal St
University Pl
Greene St
GREENWICH VILLAGE
W. Houston St
NEW MUSEUM OF CONTEMPORARY ART
Sullivan St
SO
58
57
54
53
52
51
49
48
46
45
42

Bronze

- Cottages
- Outhouses
- Dirt paths
- Isolated homes

Silver

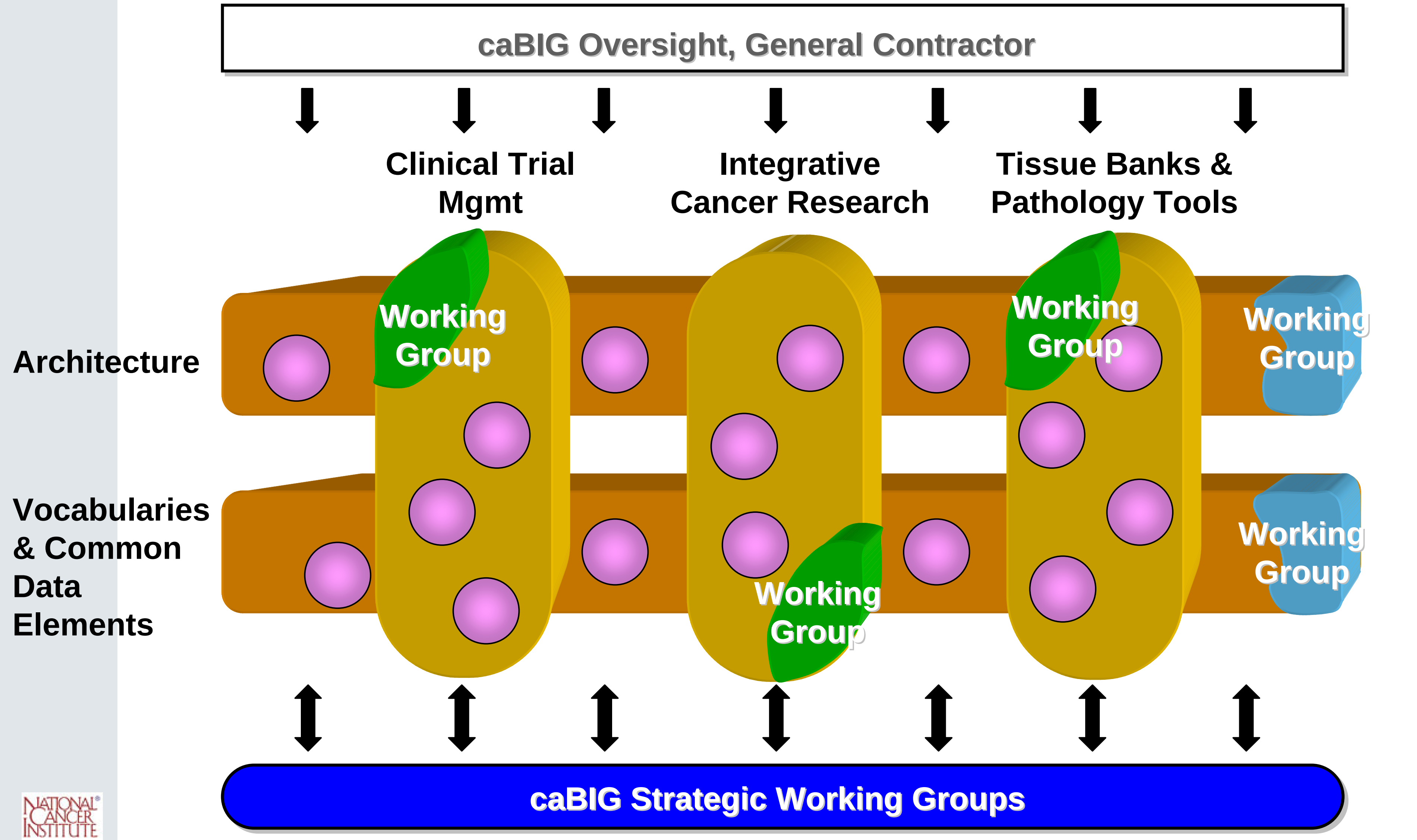
- Cottages to buildings
- Outhouses to indoor plumbing
- Dirt paths to paved streets
- Isolated homes to neighborhoods

Gold

- Village to City

caBIG cancer Biomedical Informatics Grid

Urban planning


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
Working Group
caBIG Strategic Working Groups

This land is your land...

- Open source, open access, federated
- Obligation to share, contribute to the larger biomedical community
- Value of data outlives original study

Courtesy: Charlie Mead

Interoperability

- **in·ter·op·er·a·bil·i·ty**
  - ability of a system...to use the parts or equipment of another system

    Source: Merriam-Webster web site

- **interoperability**
  - ability of two or more systems or components to exchange information and to use the information that has been exchanged.

    Source: IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries, IEEE, 1990]

Syntactic interoperability

Semantic interoperability

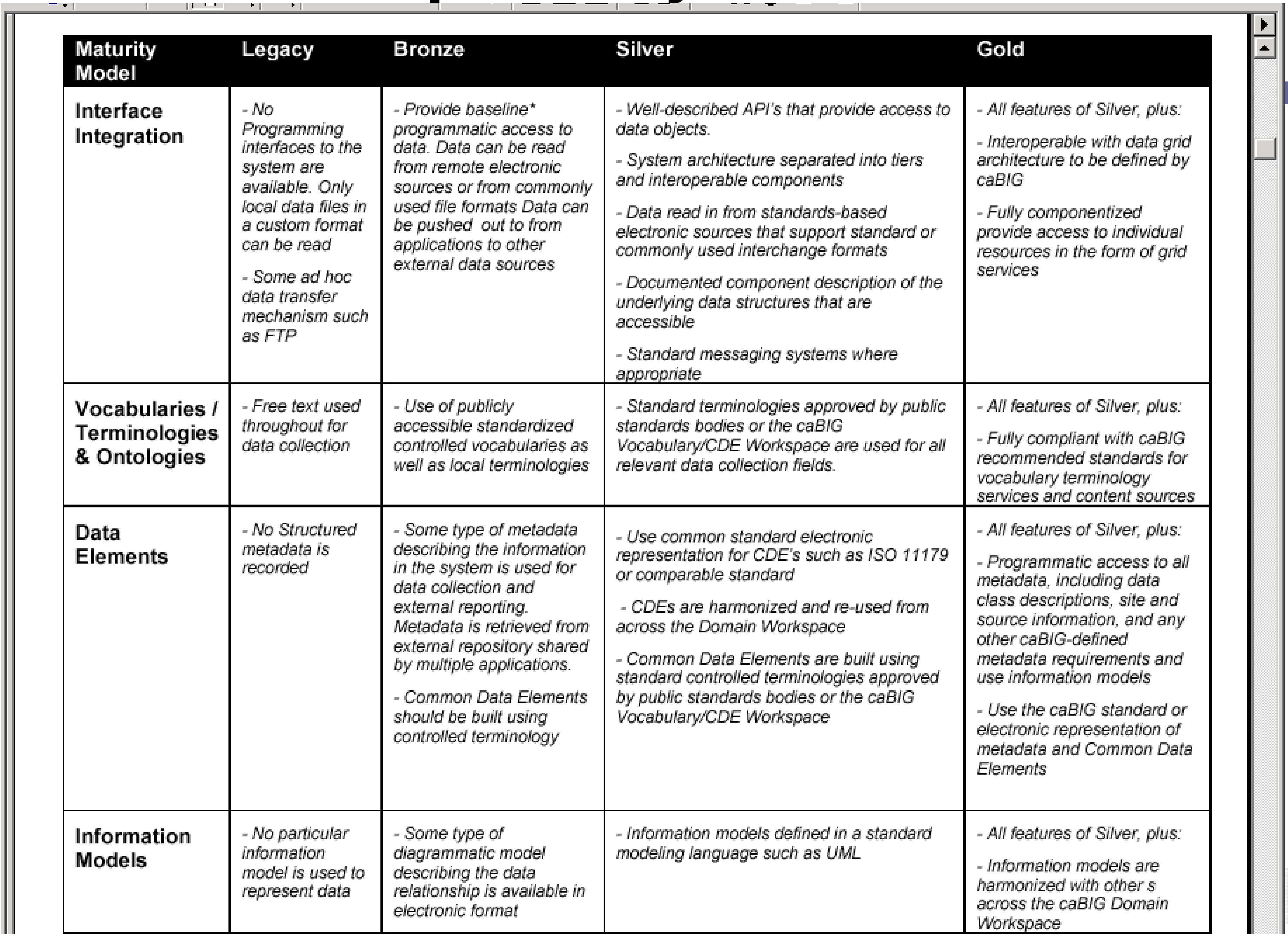
Courtesy: Charlie Mead

Pillars of Interoperability

- Common models across all domains of interest
- Foundation of rigorously defined data types
- Methodology for interfacing with controlled vocabularies

caBIG Compatibility

Levels of Compatibility



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

caBIG Silver

Information Models

- Constructed using Unified Modeling Language (UML)
- Object models expressing biomedical data classes, attributes, and relationships

Common Data Elements (CDEs)

- Metadata descriptors for cancer research data. Basis for common understanding of meaning.
- Derived from Information Models and through manual curation
- Built using standardized terminologies
- Harmonized across Domain Workspace
- Represented in standard format such as ISO/IEC 11179

Controlled Terminologies

- Standardized terminologies approved by public standards bodies or the caBIG Vocabulary-CDE workspace
- Used for all relevant data collection fields and for associated CDEs and metadata

Interfaces

- Data structures and APIs are well documented and aligned with object oriented information model
- Support for data input from standardized electronic formats and sources
- Standardized messaging interfaces where appropriate

Architecture

- Some freedom and flexibility as long as compliant with interface, model, metadata and terminology data standards
- HOWEVER, a component-based, tiered architecture is favored as a best-practice
  - Provides maximum flexibility
  - Best suited for layering an information model over a database
  - Allows problem space to be broken into manageable segments

Silver to Gold

Tiered Approach

- Silver compatible systems will be largely “Gold-ready”
- Interfaces and adaptors will bridge the tiers
- Allows for variation in types of systems that can ‘plug into’ caBIG
- Gold layer will provide standardized resource advertising, discovery and access framework for all caBIG compliant systems and tools

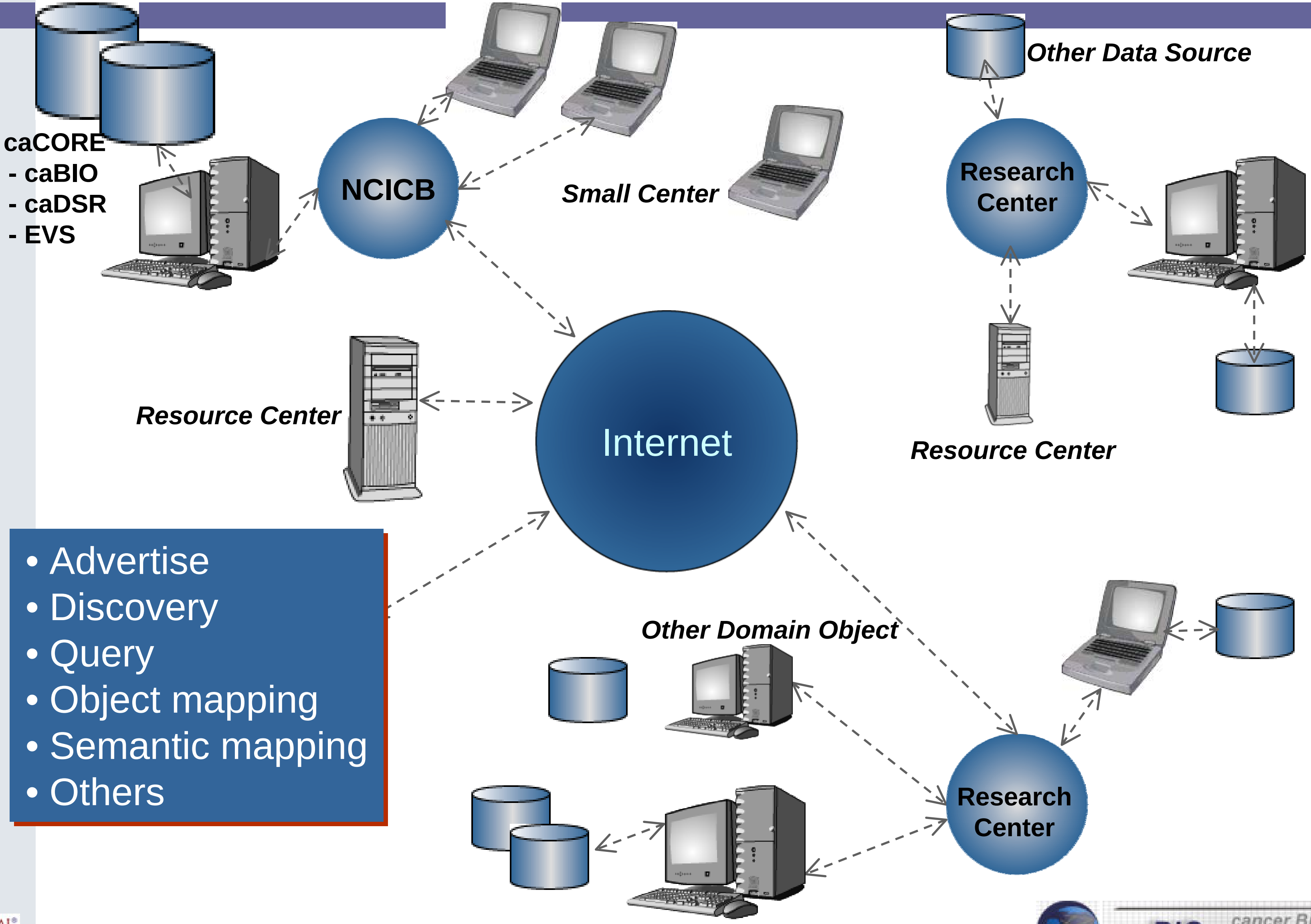
caBIG Gold: caGRID Phase I

The Yellowbrick Road to Gold

- Establish use cases
- Define requirements
- Conduct technology survey and evaluation
- Design prototype architectural model
- Develop prototype/reference implementation
- Publicize and discuss lessons learned
- Repeat…until ready for production deployment in caBIG

Gold

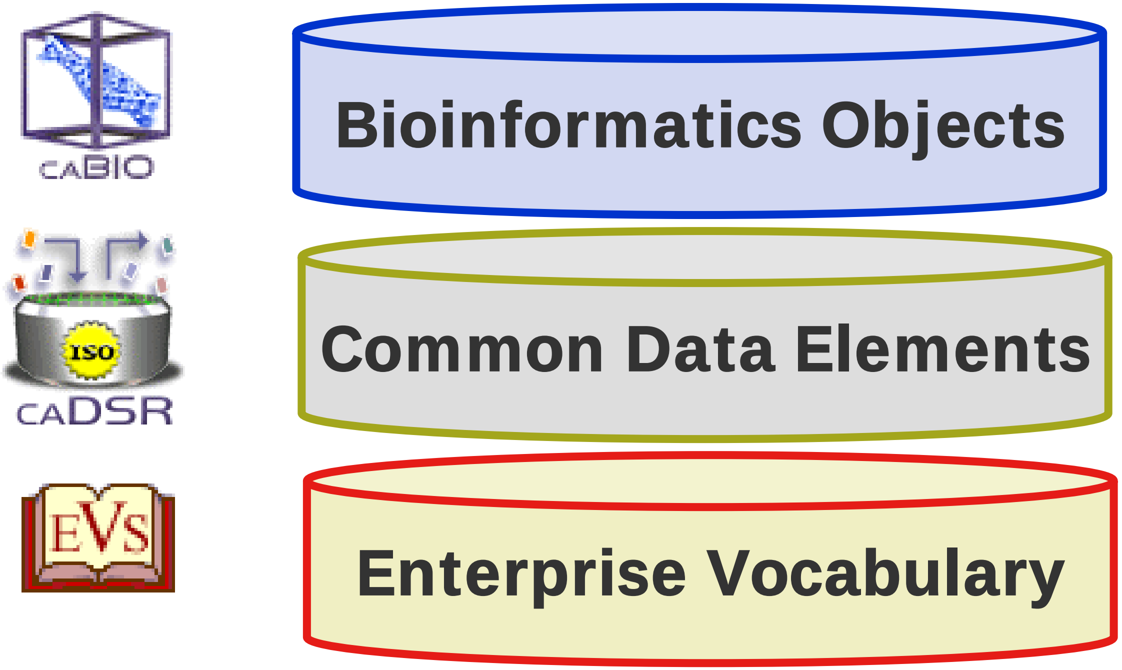
Grid requirements



- Advertise
- Discovery
- Query
- Object mapping
- Semantic mapping
- Others

Prototype layered upon caCORE

- caCORE is the ‘Silver’ technology stack developed and operated by the NCI



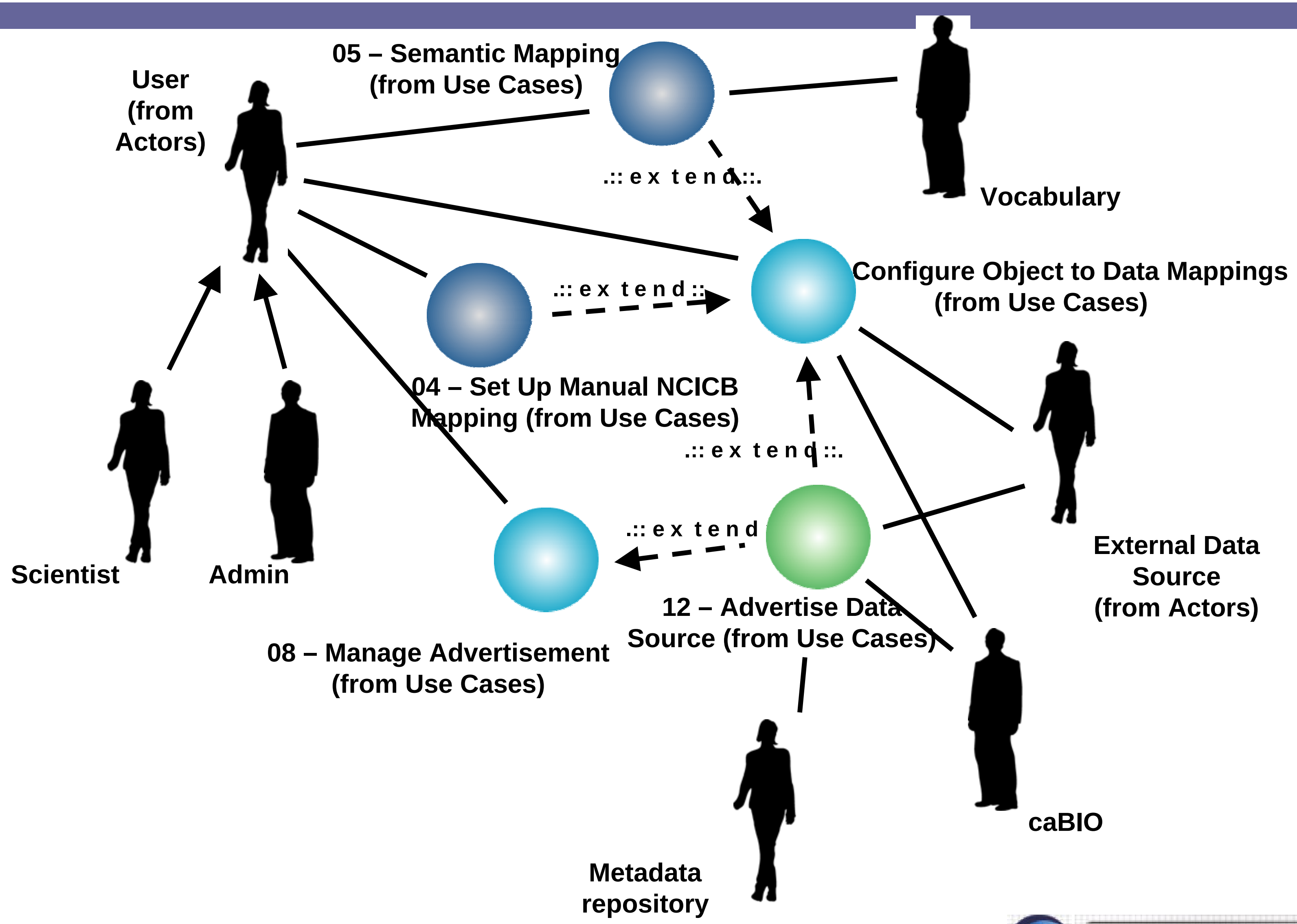
Bioinformatics Objects

Common Data Elements

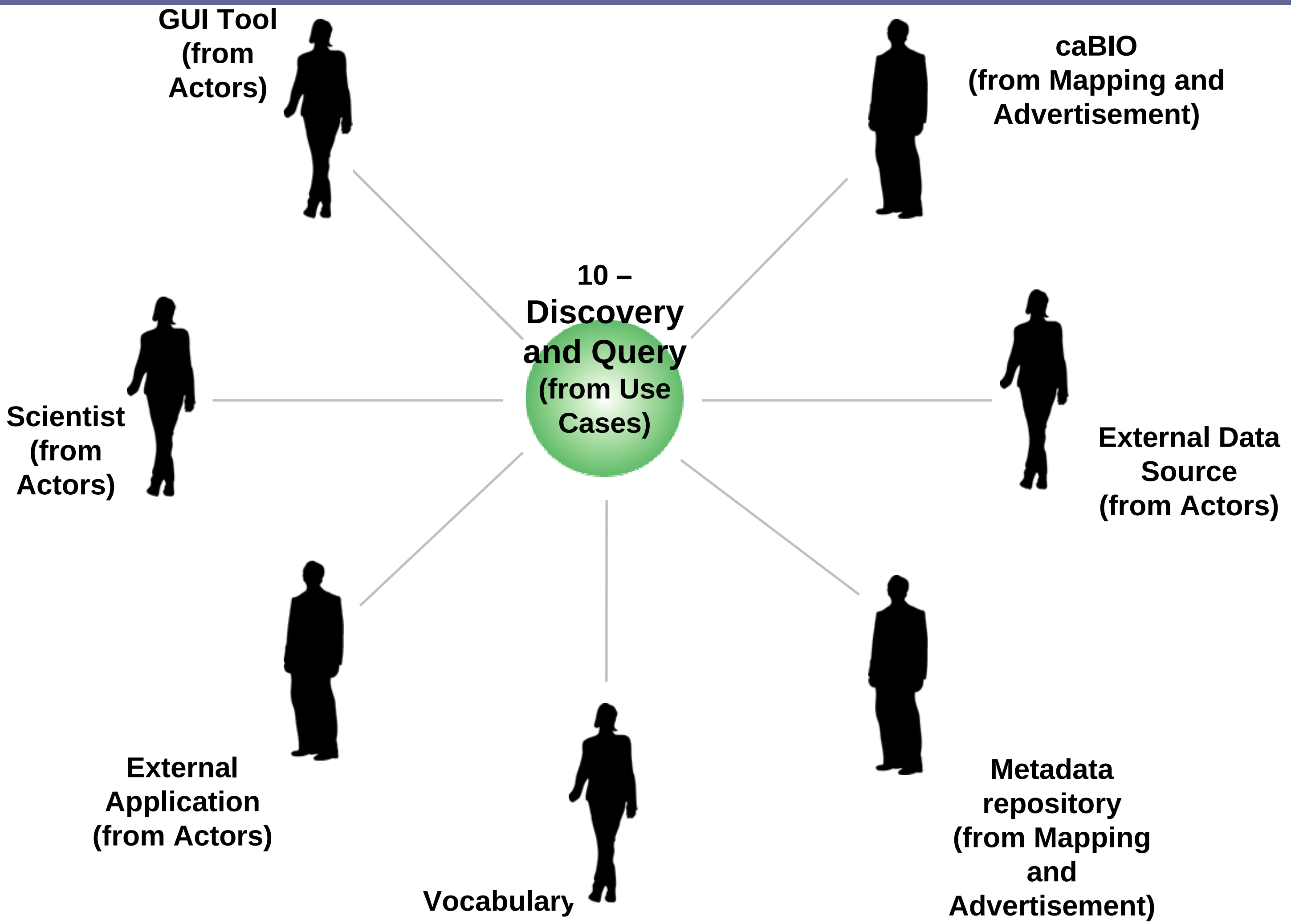
Enterprise Vocabulary

Gold

Advertisement and Mapping


05 – Semantic Mapping (from Use Cases)
User (from Actors)
.:: e x t e n d ::.
Vocabulary
.:: e x t e n d ::
Configure Object to Data Mappings (from Use Cases)
04 – Set Up Manual NCICB Mapping (from Use Cases)
.:: e x t e n d ::.
Scientist
Admin
.:: e x t e n d
External Data Source (from Actors)
12 – Advertise Data Source (from Use Cases)
08 – Manage Advertisement (from Use Cases)
caBIO
Metadata repository

Discovery and Query



GUI Tool
(from
Actors)

caBIO
(from Mapping and
Advertisement)

10 –
Discovery
and Query
(from Use
Cases)

Scientist
(from
Actors)

External Data
Source
(from Actors)

External
Application
(from Actors)

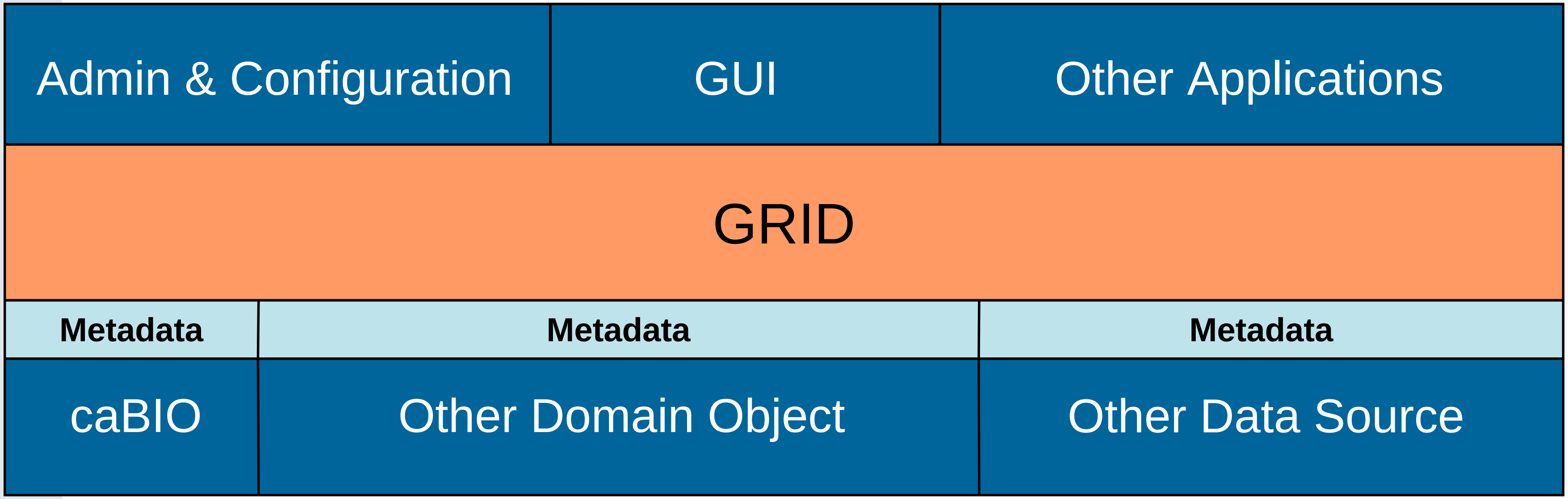
Vocabulary

Metadata
repository
(from Mapping
and
Advertisement)

Requirement Prioritization

	Prototype
Goals	-Query Data and Discovery Service – Full -Advertise services – Command line -Startup – Command line -Shutdown – Command line -Install – Basic
Requirement	-Define and prioritize requirements -Perform technology evaluation (Grid, Semantics) -Define architecture -Implement prototype – caBIO -Semantics - service level -Test GRID technology/framework

Preliminary Architecture



Admin & Configuration

GUI

Other Applications

GRID

Metadata

Metadata

Metadata

caBIO

Other Domain Object

Other Data Source

Some of the projects, technologies and standards that were examined

OGSA-DAI

(Data grid)

Data GRID

(Data grid project)

Jena2

(Semantics)

SDSC

SRB

STORAGE RESOURCE BROKER

(Data grid application)

myGrid

(Grid project)

the globus alliance

(Grid infrastructure framework)

The North Carolina BioGrid Project

Web Services

moby

(Web service registry for Bioinformatics)

BIRN

BIOMEDICAL INFORMATICS RESEARCH NETWORK

(Grid project)

Chinook: P2P Bioinformatics

Canada's Michael Smith Genome Sciences Centre

JXTA

(P2P technology)

AVAKI

(Data grid application)

OGSA and Globus Toolkit 3

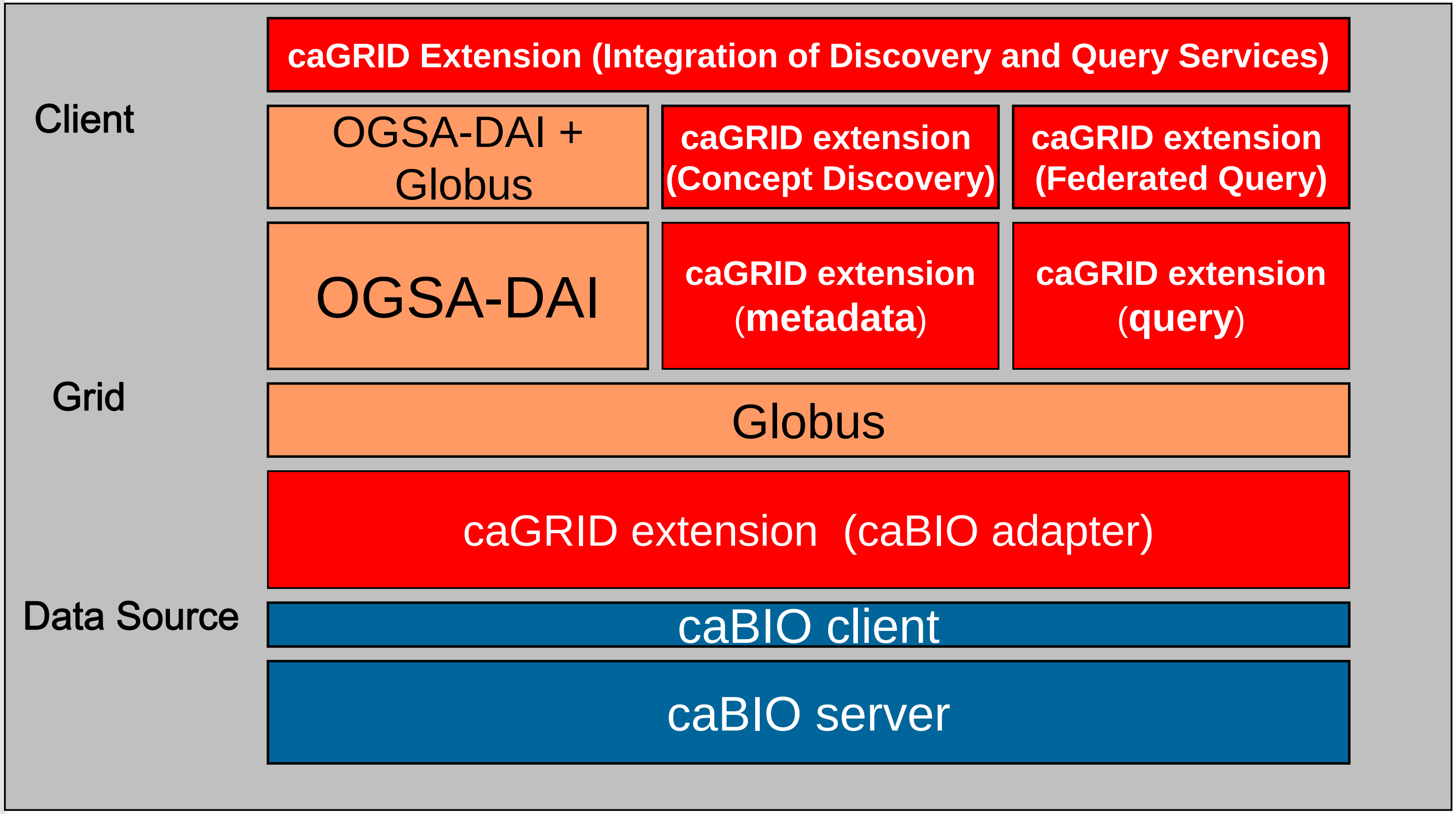
- **OGSA:** Open standard architecture for next generation grid-services
- **OGSI:** Core component of OGSA, provides a uniform way to describe grid services and defines a common pattern of behavior for all grid services.
- **GT3:** The Globus software technology toolkit version 3 is the major reference implementation of the OGSI standard.

OGSA-Data Access and Integration (OGSA-DAI)

- Middleware to assist with access and integration of data from separate data sources via the grid.
- The project was conceived by the UK Database Task Force and is working closely with the Global Grid Forum DAIS-WG and the Globus Team.

Gold

caGRID Phase I Architecture



Client

caGRID Extension (Integration of Discovery and Query Services)

OGSA-DAI + Globus

caGRID extension (Concept Discovery)

caGRID extension (Federated Query)

Grid

OGSA-DAI

caGRID extension (metadata)

caGRID extension (query)

Globus

caGRID extension (caBIO adapter)

Data Source

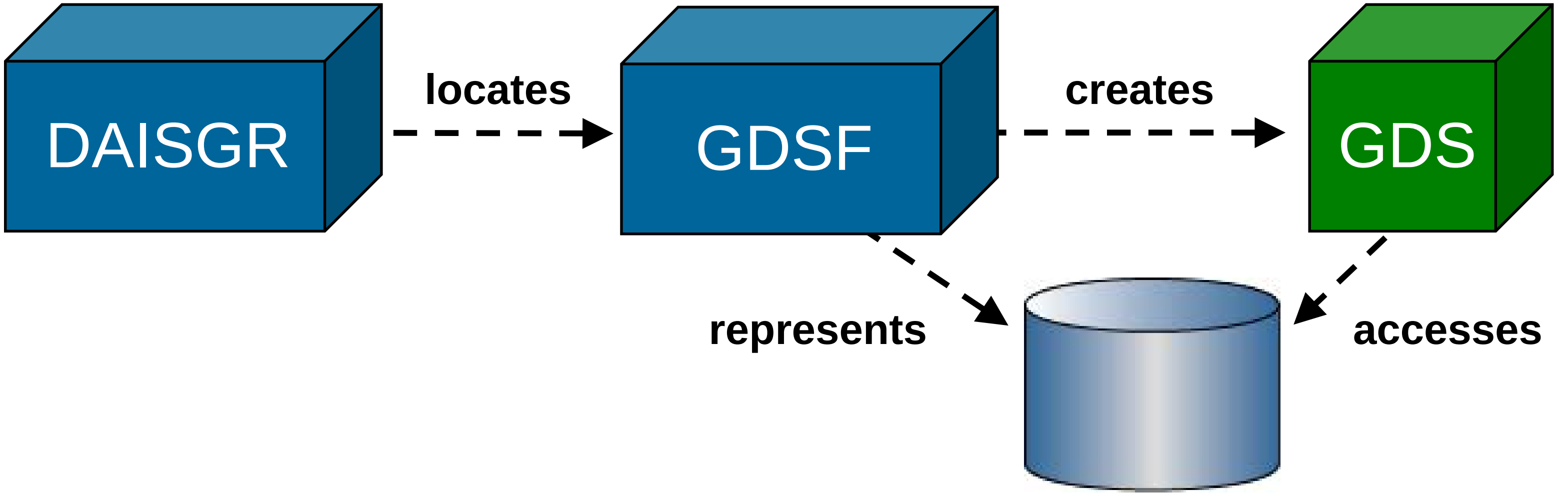
caBIO client

caBIO server

Gold

OGSA-DAI Services

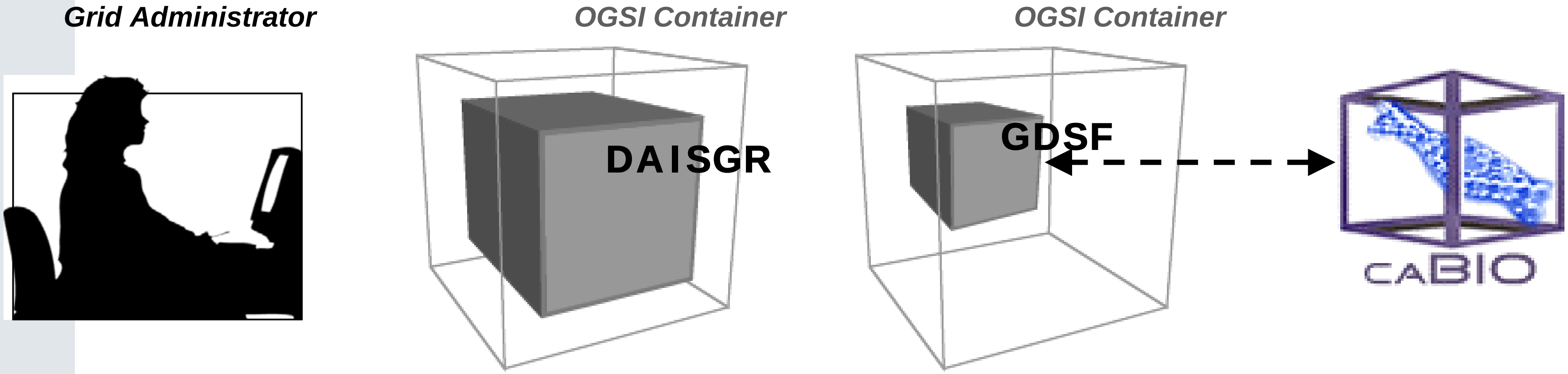
- OGSA-DAI uses three main service types
  - **D**ata **A**ccess & **I**ntegration **S**ervice **G**roup **Registry** (DAISGR) for discovery
  - **G**rid **D**ata **S**ervice **Factory** (GDSF) to represent a data resource
  - **G**rid **D**ata **Service** (GDS) to access a data resource



Data Resource

Interaction Model: Start up

1. Start OGSI containers with persistent services.
2. Here GDSF represents caBIO database.



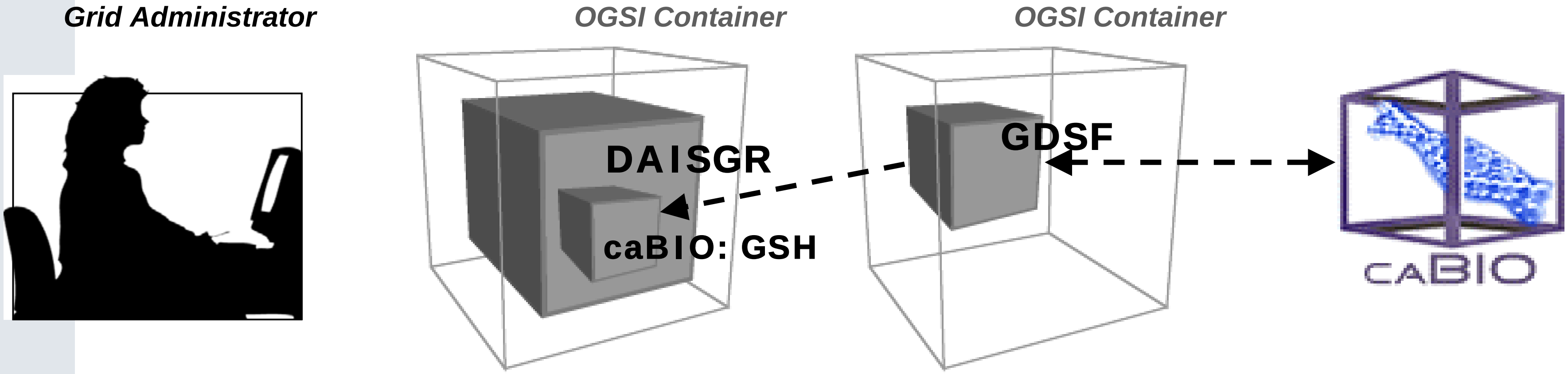
-DAISGR (registry) for discovery

-GDSF (factory) to represent a data resource

-GDS (data service) to access a data resource

Interaction Model: Registration

3. GDSF registers with DIASGR



-DAISGR (registry) for discovery

-GDSF (factory) to represent a data resource

-GDS (data service) to access a data resource

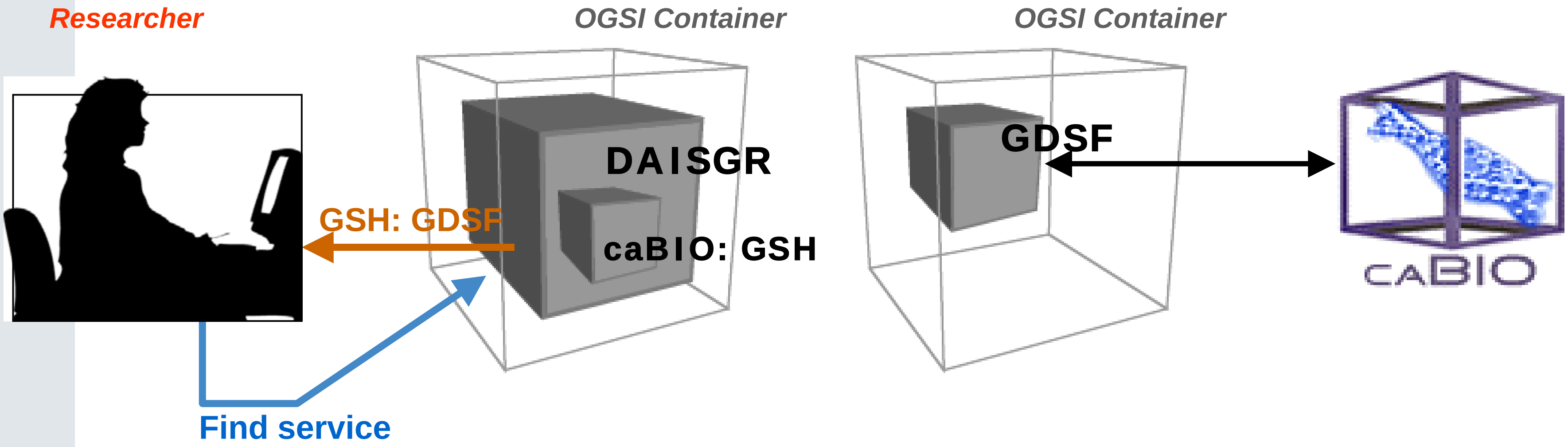
Gold

Interaction Model: Discovery

4. Client wants to know about caBIO:

(i) Query the GDSF directly if known or

(ii) Identify suitable GDSF through DAISGR.



-DAISGR (registry) for discovery

-GDSF (factory) to represent a data resource

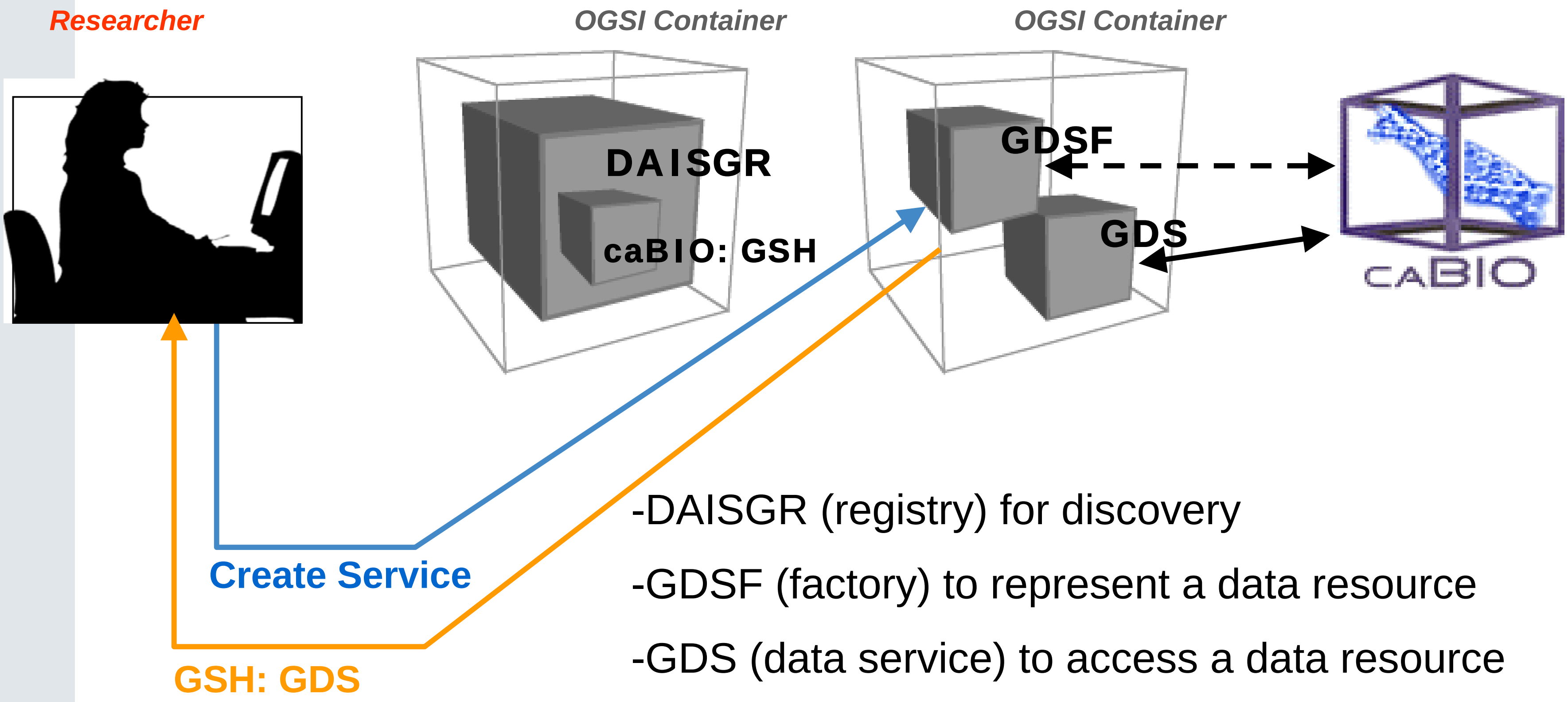
-GDS (data service) to access a data resource

-Grid Service Handler (GSH)

Gold

Interaction Model: Service Creation

5. Having identified a suitable GDSF client asks a GDS to be created.



-DAISGR (registry) for discovery

-GDSF (factory) to represent a data resource

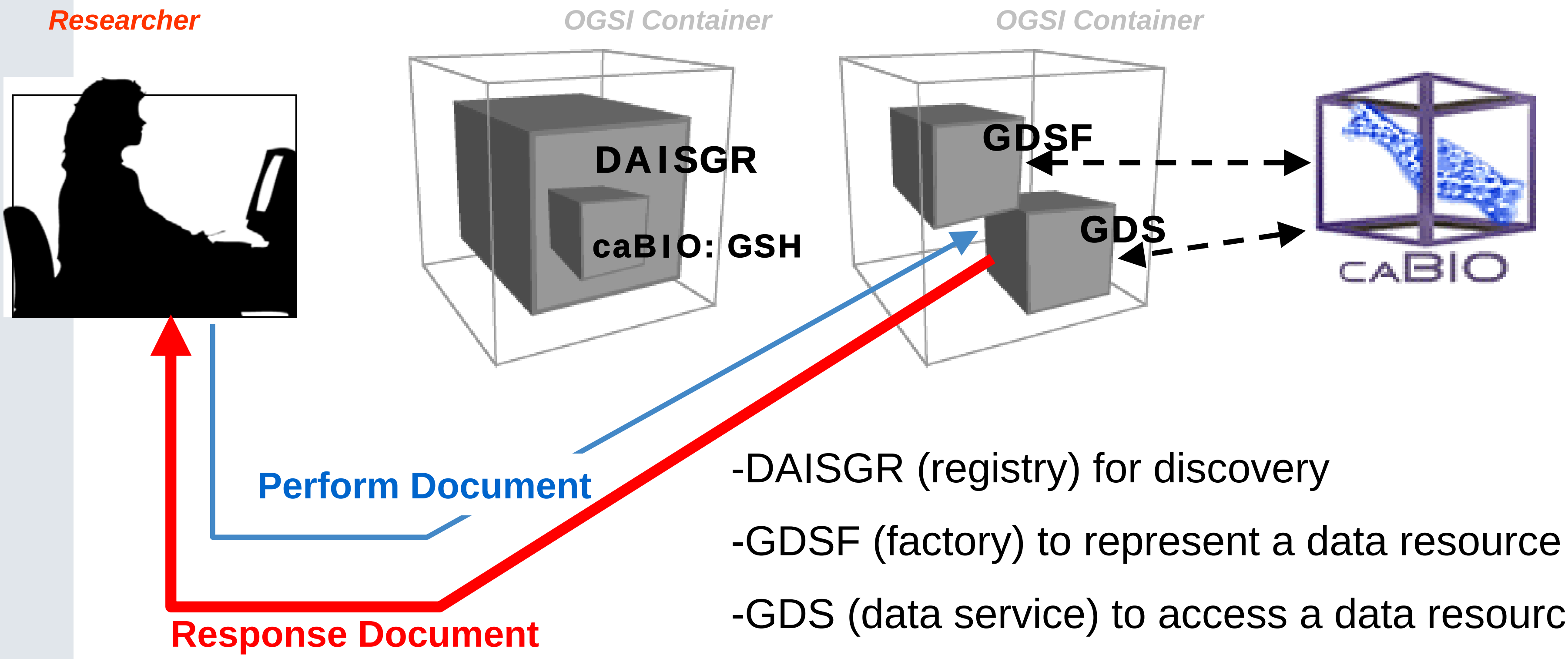
-GDS (data service) to access a data resource

-Grid Service Handler (GSH)

Gold

Interaction Model: Perform

6. Client interacts with GDS by sending Perform documents.
7. GDS responds with a Response document.
8. Client may terminate GDS when finished or let it die naturally.



-DAISGR (registry) for discovery

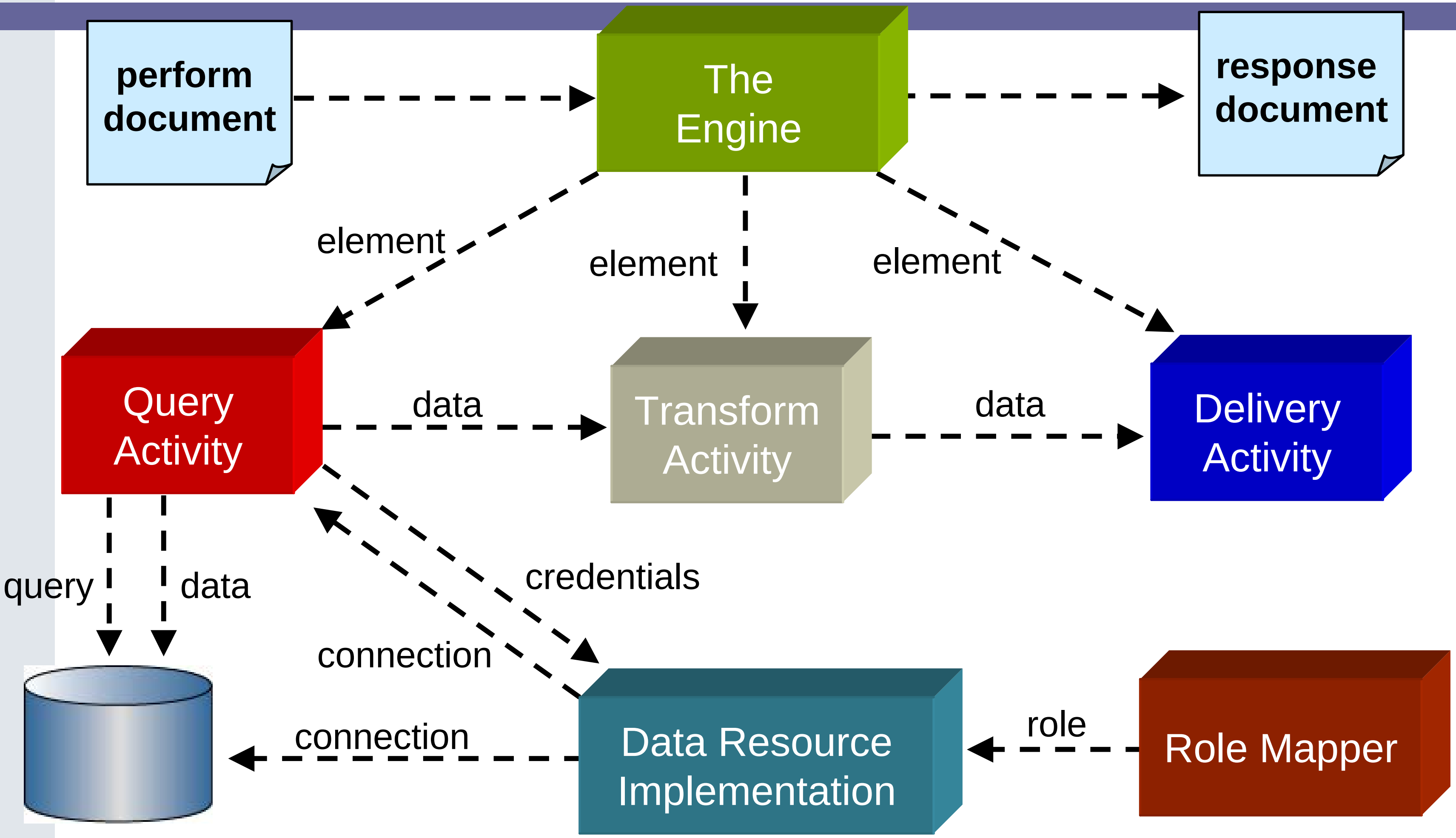
-GDSF (factory) to represent a data resource

-GDS (data service) to access a data resource

-Grid Service Handler (GSH)

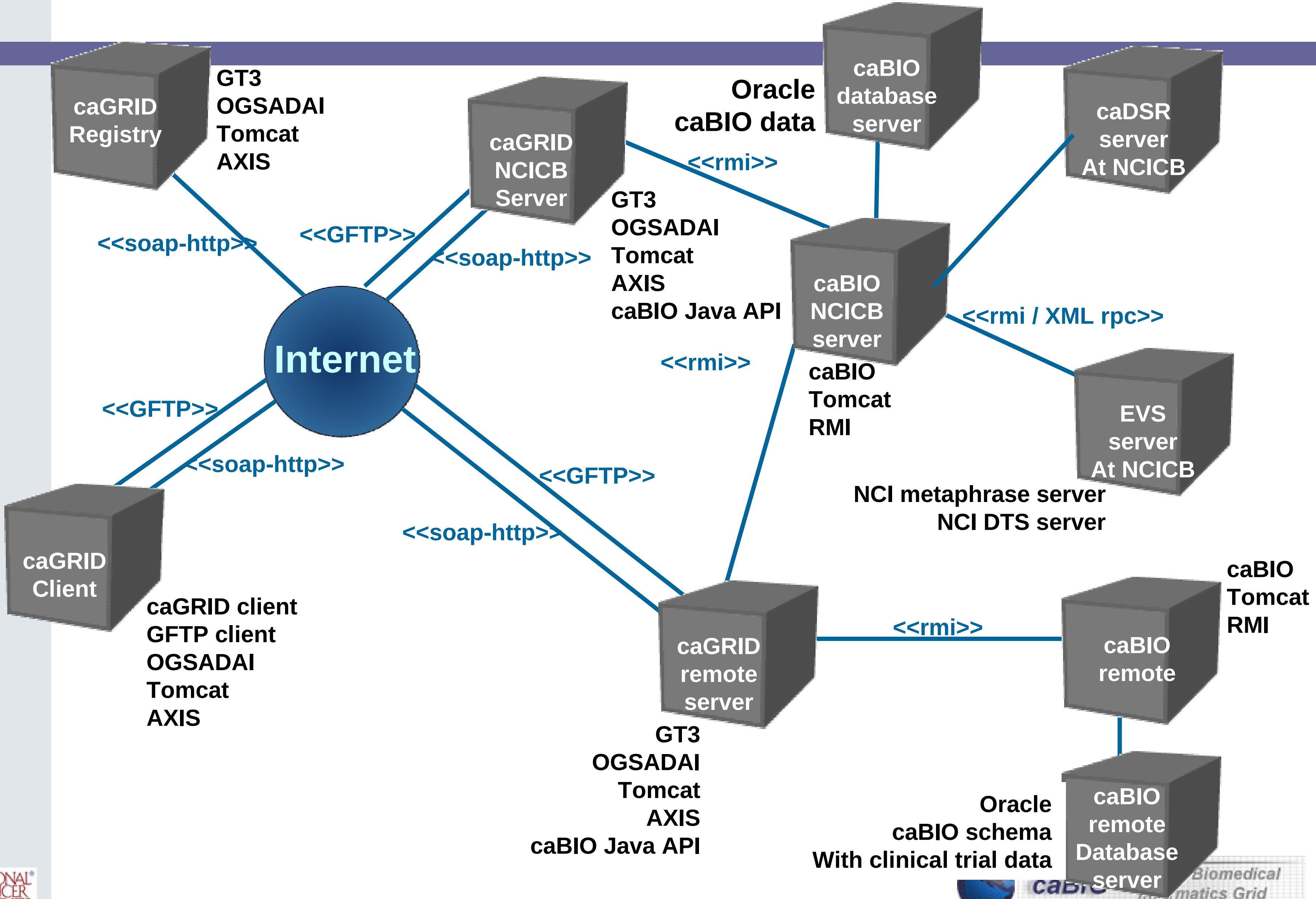
GDS Internals

Gold



Gold

caGRID Prototype Deployment



caGRID Registry

GT3
OGSADAI
Tomcat
AXIS

<<soap-http>>

caGRID NCICB Server

<<GFTP>>

<<soap-http>>

GT3
OGSADAI
Tomcat
AXIS
caBIO Java API

Oracle
caBIO data

caBIO database server

<<rmi>>

caDSR server At NCICB

caBIO NCICB server

caBIO
Tomcat
RMI

<<rmi / XML rpc>>

EVS server At NCICB

NCI metaphrase server
NCI DTS server

Internet

<<GFTP>>

<<soap-http>>

caGRID Client

caGRID client
GFTP client
OGSADAI
Tomcat
AXIS

<<GFTP>>

<<soap-http>>

<<rmi>>

caGRID remote server

GT3
OGSADAI
Tomcat
AXIS
caBIO Java API

<<rmi>>

caBIO remote

caBIO
Tomcat
RMI

Oracle
caBIO schema
With clinical trial data

caBIO remote Database server

Gold

Lesson Learned

- There is an inherent learning curve in implementing grid technologies
- Grid technologies are still maturing and preparation for upgrades is essential
- Common meta data structure and terminology is necessary to effectively describe services and data
- A common query language is important to support federated queries

caBIG &
The Mobius Project

http://www.projectmobius.org/

Scott Oster, Shannon Hastings, Stephen Langella,

Tahsin Kurc, Joel Saltz

Ohio State University

Department of Biomedical Informatics

Multiscale Computing Laboratory

Mobius Project Overview

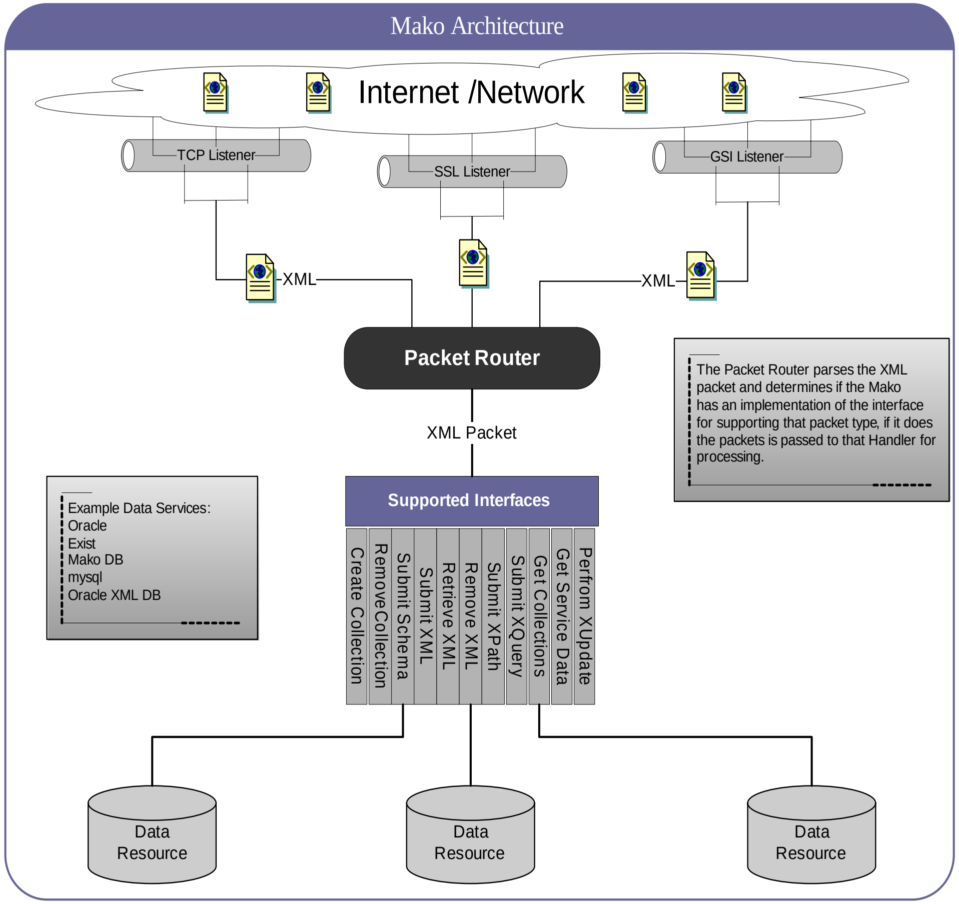
- Identifies, defines, and builds a set of services and protocols enabling the management and integration of both data and data definitions.
- Features:
  - distributed creation, versioning, management of data models and data instances
  - on demand creation of databases
  - federation of existing databases
  - querying of data in a distributed environment.
- Consists of three main components:
  - The protocol definitions.
  - The definition of service interfaces for utilizing the protocol.
  - Initial service implementation.

Mobius Services

- Mobius Core Services
  - Mako -- Federated Ad hoc Storage Services
  - GME -- Global Model Exchange
  - DTS -- Data Translation Service

Mako Service

- Exposes existing data services as XML data services through a set of well defined service interfaces based on the Mako protocol. (GGF/DAIS XML Realization Specification).
- Enables configuration file controllable binding of:
  - Network Listeners
  - Supported Interfaces
  - Protocol request implementation



The Packet Router parses the XML packet and determines if the Mako has an implementation of the interface for supporting that packet type, if it does the packets is passed to that Handler for processing.

Example Data Services:
Oracle
Exist
Mako DB
mysql
Oracle XML DB

Data Definition Management

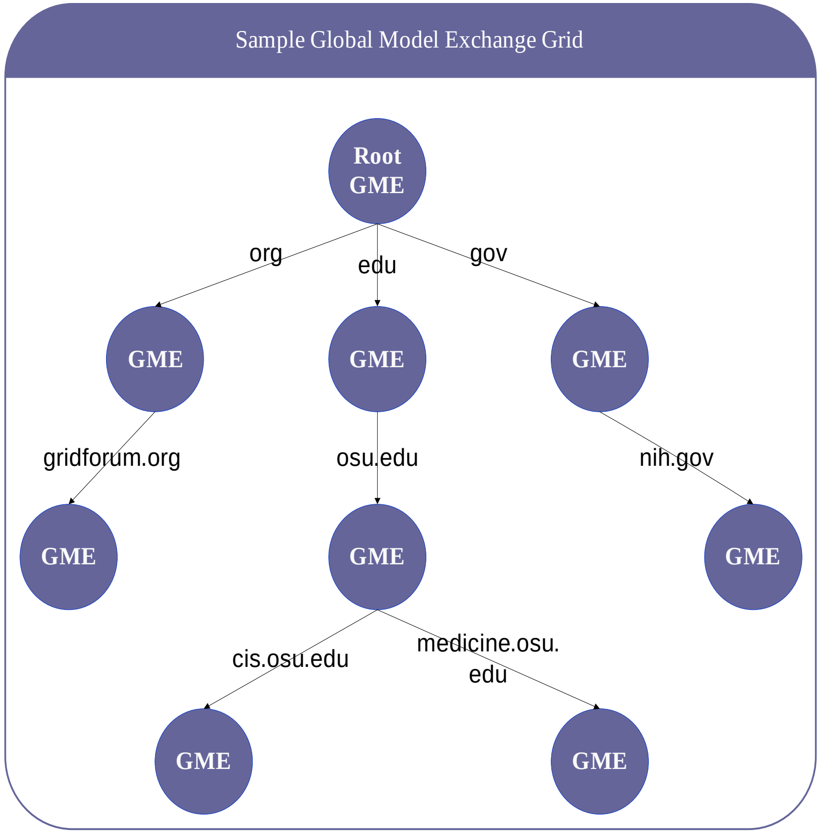
- Need for a global data definition management!
  - What is “global data definition” (Global Schema)?
  - Promote creation and evolution of standard definitions of data types.
  - For communication between multiple institutions they must agree on a common structure or a mapping between structures.
  - Allow for sharing and discovery of data definitions in a grid environment.

Global Schema Issues

- User/Organization defined entities
  - e.g.: my “person” != your “person”
- Changing schemas
- Schemas disappear
- Prevent conflicting schemas
- Discovering schemas
- Multiple definitions of similar schemas for different communities (syntactic / semantic mapping)

Global Model Exchange Service

- Manages the Global Schema
  - handles presented issues
- Provides submission and discovery protocol
- Scale
- Replicate
- Cache
- DNS like architecture
  - hierarchical parent child tree structure



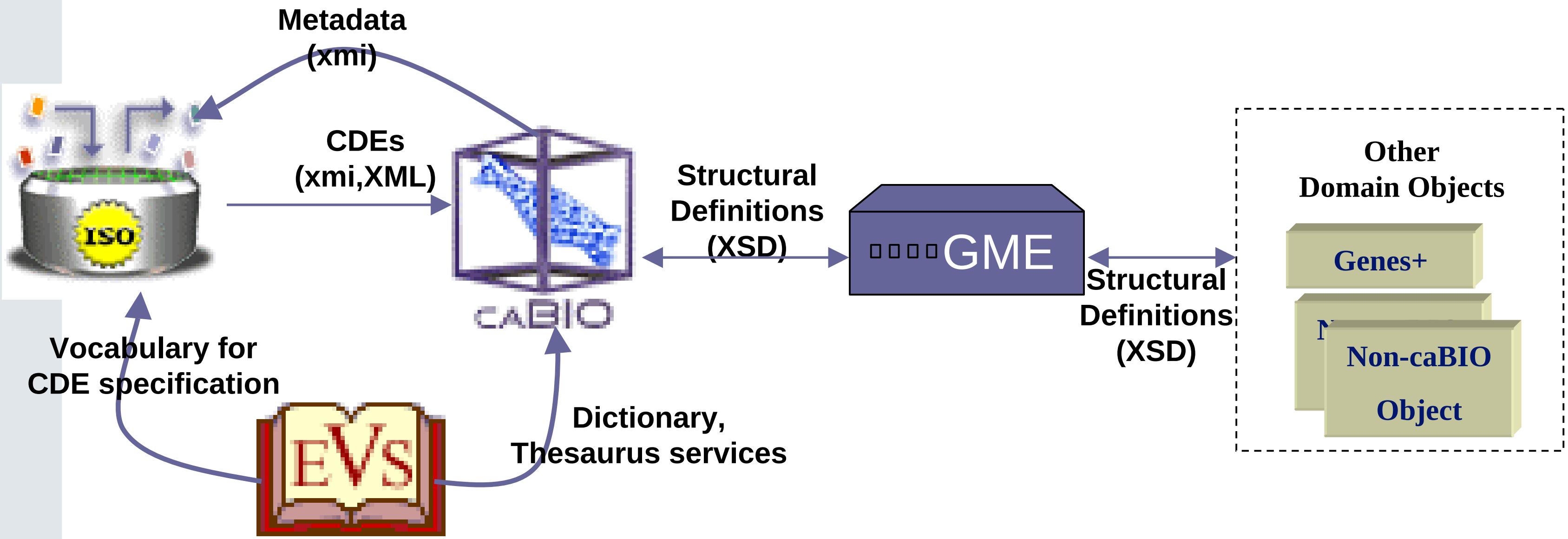
Technologies

- Protocol is XML with support for binary attachments
  - Language independent
  - Platform independent
  - Grid communication protocol independent
- Service Definitions and Initial Implementations are Java
  - Platform Independent
  - Limited C++ client API has been implemented

Potential Uses of Mobius in caBIG

Mobius in caBIG: GME

- Use Cases:
  - caBIO Object Managers validate Domain Objects against schemas in GME
  - caBIO and non-caBIO clients publish schemas to GME and create data which validates against them
  - Institutions are able to communicate about caBIO objects, extensions to caBIO objects, and objects not present in caBIO using the same mechanism



Mobius in caBIG: Mako

- Utilize Mako service to virtualize data services
  - Expose data sources to caBIG Grid using Mako service

Mobius in caBIG: MakoDB

- Provide data cache utilizing Mako and MakoDB
  - Service interaction/collaboration for computation may require storage of temporary results and/or data cache
  - Utilize Mako's ability to generate on demand databases from schemas
  - Used locally by clients or as a Grid Service

caGRID Phase II

caGRID Phase II

- Capture Use Cases from caBIG Domain Workspaces
- Evaluate caGRID, Mobius and other technologies to select best components
- Establish target architectural design, review with caBIG Architecture Workspace
- Conduct a second round of prototyping and reference implementations with caBIG Domain Workspaces. Adjust designs as needed
- Plan for production implementation in caBIG beginning mid-2005.

Decisions to date

- XML will be primary interchange format for grid services
- XML Schema plus additional semantic metadata will describe structure and semantics of grid data
- Data services will represent data as objects, from UML information models, as serialized XML
- Globus Toolkit 3.2 and OGSA-DAI 4.0 will be basis for next prototypes. Planning for Globus TK 4.0, which implements a new grid standards, will be included

Issues to be tackled

- Universal data object identifiers, needed to satisfy computational biology use cases
  - Exploring Life Science Identifiers (LSIDs), ISO OIDs, and other potential solutions
- Caching for large result sets and performance requirements
- Standard query language needed to interrogate all grid services
- Representation of data analytical services in the grid
- Authentication/Authorization infrastructure
  - Use cases still not clear, but expected to be important

Acknowledgements

caBIG Architecture WS

Fred Hutchinson
Ohio State
Duke
Cold Spring Harbor Labs
Fox Chase
Siteman/Wash. U.
Holden/U. Iowa
U. Pittsburgh
Lombardi/Georgetown
Mem. Sloan Kettering
U. Chicago
Oregon Health Science
NCI Center for Cancer Research
NCI Center for Bioinformatics

caGRID Phase I

William Sanchez
Manav Kher
Brian Gilman
Steve Lagou
SAIC
Panther Informatics
Booz Allen Hamilton
Terrapin Systems

OSU - Mobius

Joel Saltz
Scott Oster
Shannon Hastings
Stephen Langella
Tahsin Kurc

NATIONAL CANCER INSTITUTE

caBIG cancer Biomedical Informatics Grid

Links to more information

- NCICB

  http://ncicb.nci.nih.gov

- caBIG

  http://cabig.nci.nih.gov

  caGRID and Mobius documents posted on Architecture Workspace page