



Architecture Requirements and Use Case Drivers of caGrid

Shannon Hastings (hastings@bmi.osu.edu)

Scott Oster (oster@bmi.osu.edu)

Stephen Langella (langella@bmi.osu.edu)

Department of Biomedical Informatics

Multiscale Computing Laboratory

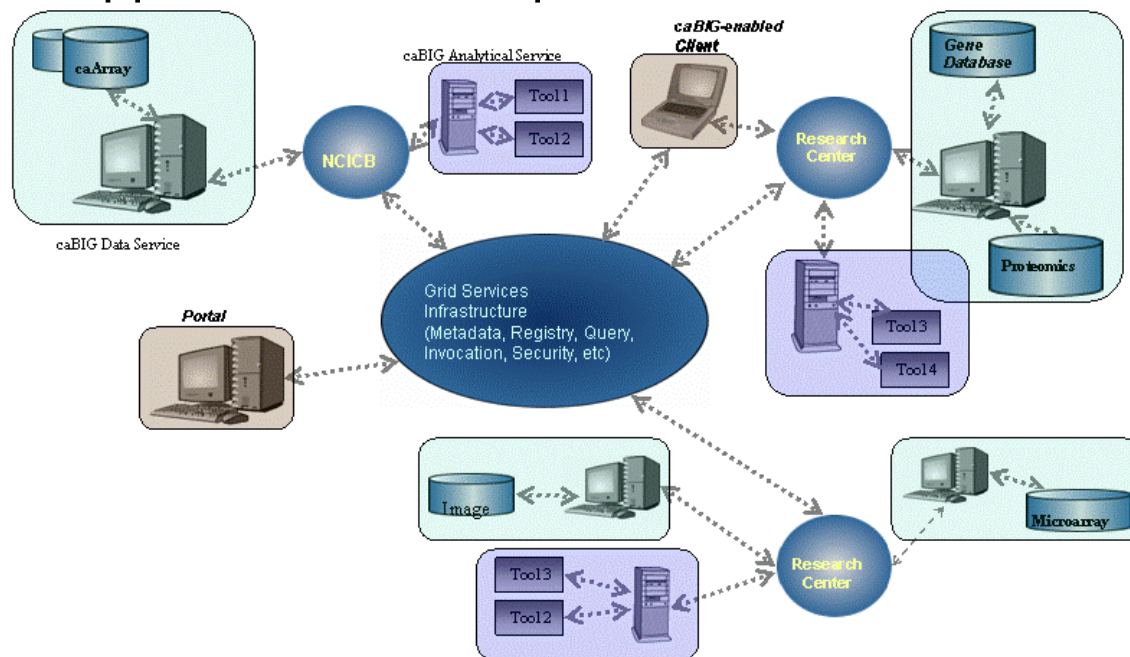
Ohio State University

Outline

- ▶ caBIG/caGrid Overview
- ▶ Roadmap
- ▶ Use Case Drivers and Requirements
- ▶ Interoperable
 - Syntax and Semantics
- ▶ Model Driven
- ▶ Semantically Discoverable
- ▶ Secure and Manageable
 - Multi-institution
 - Regulatory
- ▶ Conclusions

caBIG Overview

- ▶ The **cancer Biomedical Informatics Grid**, or **caBIG™**, is a voluntary network or grid connecting individuals and institutions to enable the sharing of data and tools, creating a World Wide Web of cancer research. The goal is to speed the delivery of innovative approaches for the prevention and treatment of cancer.



caGrid Overview

- ▶ Driven primarily from the scientific use cases identified from the domain workspaces (community) of caBIG, caGrid provides the **core enabling infrastructure** necessary to compose the Grid of caBIG.
- ▶ caGrid is a service-oriented architecture and provides the implementation of:
 - the required core services
 - toolkits and wizards for the development and deployment of community provided services
 - APIs for building client applications
 - sample client applications for interacting with the current test bed installation.

Current caGrid Roadmap

- ▶ caGrid 0.5 released August 2005
 - Leveraged Technologies
 - Globus Toolkit 3.2
 - OGSA-DAI 5.0
 - Multi-site test bed consisting of data and analytical service reference implementations

- ▶ caGrid 1.0 coming mid to late 2006
 - Key planned extensions
 - possible move to Globus Toolkit 4.0 as base toolkit
 - security architecture extensions
 - workflow support
 - analytical service toolkit extensions
 - distributed query framework



Use Case Example

- ▶ **caBIG Workspace:** Tissue Banks and Pathology Tools
- ▶ **Scenario:** A research scientist learns about the caBIG caTissue biospecimen informatics service. He would like to 'log on' to the system to search for rare cases of pilocytic astrocytomas in patients with documented Neurofibromatosis where frozen tissue is available for gene expression analysis. He will eventually request samples of tissues that he finds for a (potentially collaborative) research project. What assurances does he need to provide (e.g. IRB approval, certificate of HIPAA training, Code Access agreement) and to whom?
- ▶ **Problem Abstract:** This scenario poses the question about data types, semantic discovery, service invocation, and security. This scenario also poses a more interesting question about security as it asks what IRB / HIPAA certification is needed by an actor who is logging on to the caTissue system in a *Scientist* role (to view de-identified data) and to what entity this certification needs to be presented. A related problem is how often does this documentation need to be recertified (i.e. with every log-in, yearly, never)?
- ▶ **Points for Consideration:**
 - How does the caBIG caTissue service get onto the NCI grid?
 - How does the researcher learn about the data types caTissue is providing?
 - Who can see de-identified data in caTissue?
 - What certifications must they present?
 - Who validates these certifications?
 - How often do certifications need to be renewed?

Courtesy of: Mark A. Watson, Washington University

caGrid Architecture Use Case Drivers and Requirements

- ▶ caBIG requires that all data be strongly typed, have publicly available definitions, and have those definitions semantically harmonized.
- ▶ The introduction of caBIG will bring together numerous data sources, each with many different data types, some of which are overlapping
- ▶ Data types will evolve over time, and Grid services may require different versions of those data types
- ▶ Clients and Grid services must be able to enforce their data representations are compatible when they communicate
- ▶ Clients and Grid services may wish to leverage semantically similar yet syntactically different data types
- ▶ As a large amount of the data relates either directly or indirectly to patients, many data types have a number of security and privacy requirements
- ▶ As there will be many institutions involved local/institutional security vs. grid security policies will have to be managed.



Common Requirement Themes

- ▶ Interoperable
- ▶ Model Driven
- ▶ Semantically discoverable
- ▶ Secure and manageable

Interoperable

- ▶ Compatibility Standards defined and measured (Legacy, Bronze, Silver, Gold) for:
 - Programming and Messaging Interfaces
 - Vocabularies and Ontologies
 - Common Data Elements
 - Information Models

Interoperable - Semantic Interoperability

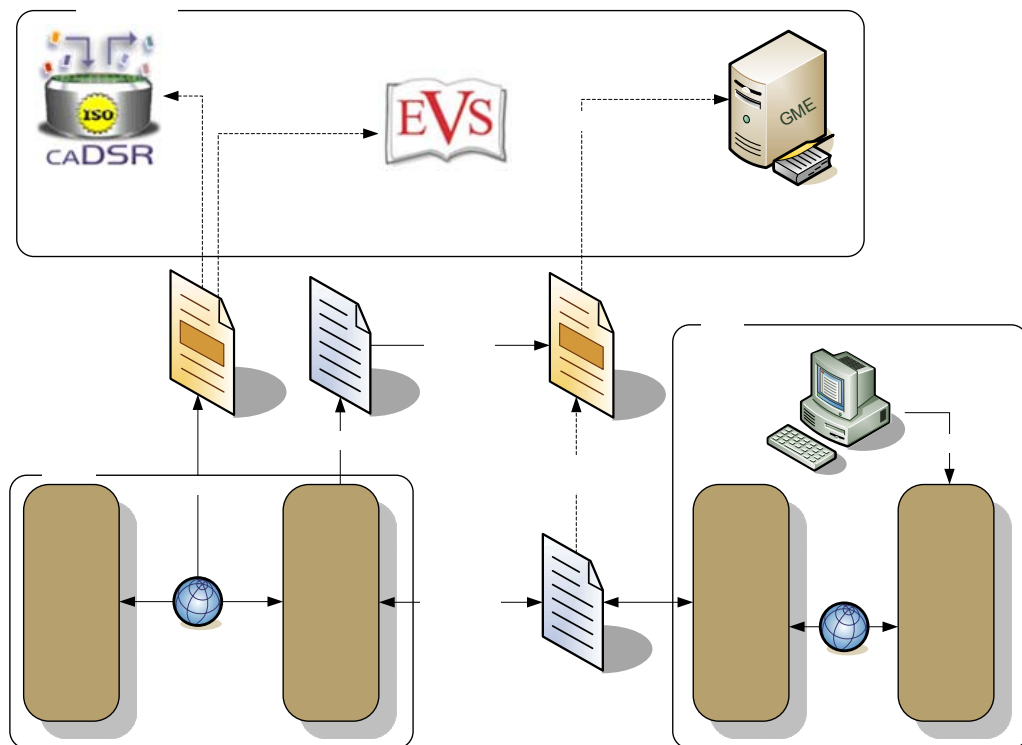
- ▶ caBIG requires that grid users provide semantic information about any data types that they are going to use in the caGrid.
- ▶ **Cancer Data Standards Repository (caDSR)**
 - NCI has a broad initiative to standardize the metadata used for cancer research. These Common Data Elements (CDEs) are developed by various NCI-sponsored organizations, then centrally stored and managed the caDSR
- ▶ **Enterprise Vocabulary Services (EVS)**
 - EVS is set of services and resources that address the need for controlled vocabulary
 - **Thesaurus**: a biomedical thesaurus
 - **Metathesaurus**: based on NLM's Unified Medical Language System Metathesaurus supplemented with additional cancer-centric vocabulary

Interoperable - Syntactic Interoperability

- ▶ caBIG requires that all data types traveling through the grid have data definitions and those definitions need to be able to be related to the semantic information about the data model.
- ▶ **Global Model Exchange (GME)**
 - GME is a DNS-like data definition registry and exchange service that is responsible for storing and linking together data models in the form of XML schema.

Model Driven

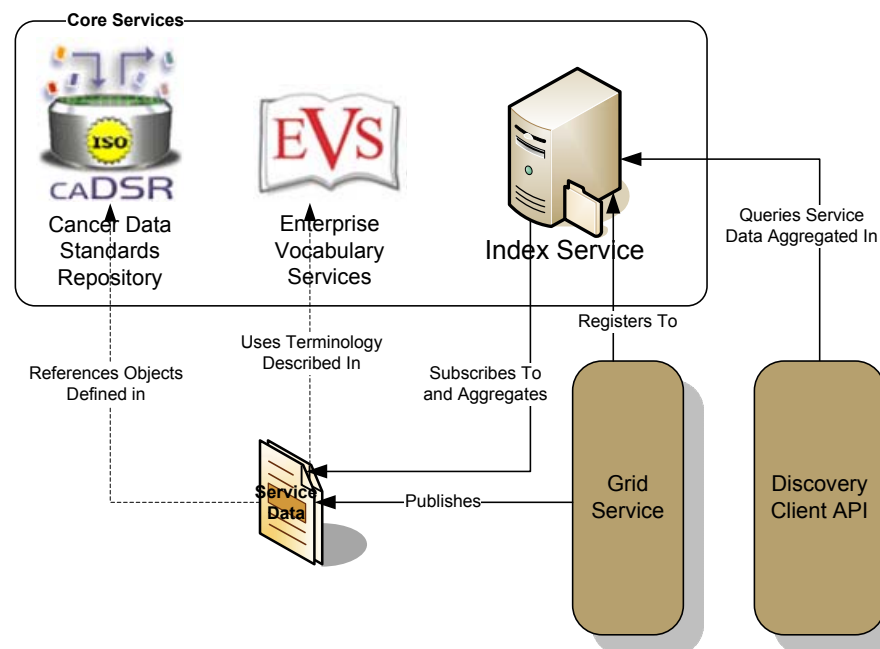
- ▶ Client and service APIs are object oriented, and operate over well-defined and curated data types
- ▶ Objects are defined in UML and converted into ISO/IEC 11179 Administered Components, which are in turn registered in the Cancer Data Standards Repository (caDSR)
- ▶ Object definitions draw from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



Semantically Discoverable

- ▶ Service Metadata (SDEs) are standardized as XML schemas stored in GME. (Common, Data Service, and Analytical Service)
- ▶ Common Metadata describes generic information about service providing Cancer Center
- ▶ Data Service Metadata describes the data exposed using terminology and objects from caDSR/EVS
- ▶ Analytical Service Metadata describes the supported operations and their inputs and outputs using terminology and objects from caDSR/EVS

- ▶ Leveraging semantic information in EVS (from which SDEs are drawn), services can be discovered by the semantics of their data types



Secure and Manageable

- ▶ Security is an especially important component of caBIG both for protecting intellectual property and ensuring protection and privacy of patient related and sensitive information.
- ▶ When security is implemented in a multi-institutional environment, such as caBIG, a challenging problem is to facilitate the management of users and user attributes.
- ▶ Furthermore, it is important to be able to leverage existing local systems for authenticating and authorizing users.

Secure and Manageable – Security Requirements

▶ **Secure Communication**

- *Authentication* - Parties involved can be assured of one another identity
- *Message Integrity* – Message sent by either party is guaranteed to same message when it is received.
- *Privacy* – Communication between the two parties can only be interpreted by the two parties

▶ **Single Sign On**

- Users and Grid Services should have one method of authenticating themselves to the grid, all services in the grid should accept this method.

▶ **Access Control on caBIG Services**

- caBIG services should be able to determine which users or services may access them

▶ **User/Organizational Attribute Management**

- Services should have a method for determining the attributes of a requesting party. Such attributes may be needed to service the request, for example a username and password is needed to perform a query on a relational database on the party's behalf.
- Attributes should be standardized such that they may be used across institutional and application boundaries

▶ **Delegation**

- caBIG services, should be able to interact with other caBIG services on a user's behalf

▶ **User/Organization Management**



caBIG

cancer Biomedical
Informatics Grid



NATIONAL
CANCER
INSTITUTE



Use Case Example

- ▶ **caBIG Workspace:** Tissue Banks and Pathology Tools
- ▶ **Scenario:** A research scientist learns about the caBIG caTissue biospecimen informatics service. He would like to 'log on' to the system to search for rare cases of pilocytic astrocytomas in patients with documented Neurofibromatosis where frozen tissue is available for gene expression analysis. He will eventually request samples of tissues that he finds for a (potentially collaborative) research project. What assurances does he need to provide (e.g. IRB approval, certificate of HIPAA training, Code Access agreement) and to whom?
- ▶ **Problem Abstract:** This scenario poses the question about data types, semantic discovery, service invocation, and security. This scenario also poses a more interesting question about security as it asks what IRB / HIPAA certification is needed by an actor who is logging on to the caTissue system in a *Scientist* role (to view de-identified data) and to what entity this certification needs to be presented. A related problem is how often does this documentation need to be recertified (i.e. with every log-in, yearly, never)?
- ▶ **Points for Consideration:**
 - How does the caBIG caTissue service get onto the NCI grid?
 - How does the researcher learn about the data types caTissue is providing?
 - Who can see de-identified data in caTissue?
 - What certifications must they present?
 - Who validates these certifications?
 - How often do certifications need to be renewed?

Courtesy of: Mark A. Watson, Washington University



caBIG

cancer Biomedical
Informatics Grid



NATIONAL
CANCER
INSTITUTE



More information

- ▶ caBIG website:
 - <https://cabig.nci.nih.gov/>
- ▶ caBIG Architecture:
 - <https://cabig.nci.nih.gov/workspaces/Architecture/>
- ▶ BMI's involvement in caBIG:
 - http://bmi.osu.edu/areas_and_projects/caBIG.cfm
- ▶ BMI's website:
 - <http://bmi.osu.edu/>