caGrid 0.5 Overview

The cancer Biomedical Informatics Grid, or caBIG™, is a voluntary virtual informatics infrastructure that connects data, utilizes research tools, and enables scientists and organizations to leverage their combined strengths and expertise in an open environment with common standards and shared tools. The current test bed architecture of caBIG, dubbed caGrid, is described in this document with respect to its technical architecture. The software embodiment and corresponding documentation of this architecture constitute the caGrid 0.5 release.

Driven primarily from the scientific use cases identified from the domain workspaces 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, and some sample client applications for interacting with the current test bed installation. A conceptual overview of these components is shown below in Figure 1.

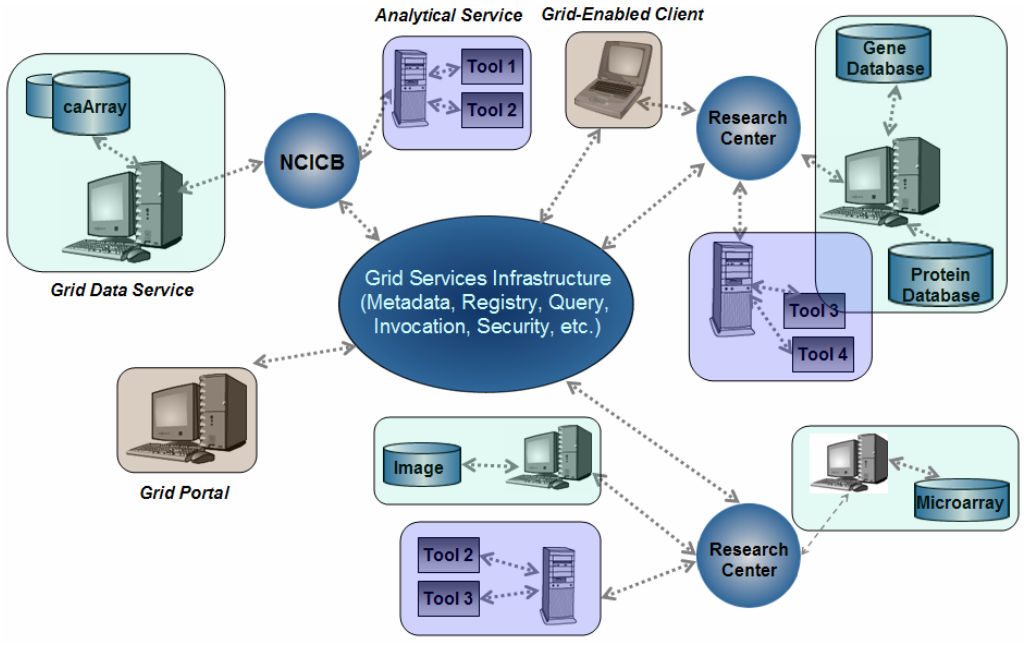

Figure 1 Overview of caGrid

Standards Compliant

As a primary principle of caBIG is open standards, caGrid is built upon the community-driven standards of Open Grid Services Architecture (OGSA). OGSA is the current specification for grid infrastructure and has been developed over the past several years in the Global Grid Forum (GGF), the community of users, developers, and vendors leading the global standardization effort for grid computing. The current release of caGrid is built using the Globus Toolkit 3.2 (GT3.2) and the OGSA Data Access Integration (OGSA-DAI) framework. GT3.2 is a reference implementation of OGSA for service creation, deployment, and invocation and provides core services such as the Index Service for service registration and discovery and Globus Security Infrastructure (GSI) for security. The OGSA-DAI framework is an implementation of OGSA for data services. It provides a set of interfaces and runtime support for implementing and deploying data sources as Grid services.

Model Driven

Extending beyond the basic grid infrastructure, caBIG specializes these technologies to better support the needs of the cancer research community. A primary distinction between basic grid infrastructure and the requirements identified and implemented in caGrid is the attention given to data modeling and semantics. caBIG adopts a model-driven architecture best practice and requires that all data types used on the grid are formally described, curated, and semantically harmonized. These efforts result in the identification of common data elements, controlled vocabularies, and object-based abstractions for all cancer research domains. caGrid leverages existing NCI data modeling infrastructure to manage, curate, and employ these data models. Data types are defined in caCORE UML and converted into ISO/IEC 11179 Administered

Components, which are in turn registered in the Cancer Data Standards Repository (caDSR). The definitions draw from vocabulary registered in the Enterprise Vocabulary Services (EVS), and their relationships are thus semantically described.

In caGrid, both the client and service APIs are object oriented, and operate over well-defined and curated data types. Clients and services communicate through the grid using Globus grid clients and service infrastructure, respectively. The grid communication protocol is XML, and thus the client and service APIs must transform the transferred objects to and from XML. This XML serialization of caGrid objects is restricted, as each object that travels on the grid must do so as XML which adheres to an XML schema registered in the Global Model Exchange (GME). As the caDSR and EVS define the properties, relationships, and semantics of caBIG data types, the GME defines the syntax of the XML serialization of them. Furthermore, Globus services are defined by the Web Service Description Language (WSDL). The WSDL describes the various operations the service provides to the grid. The inputs and outputs of the operations, among other things, in WSDL are defined by XML schemas. As caBIG requires that the inputs and outputs of service operations use only registered objects, these input and output data types are defined by the XSDs which are registered in GME. In this way, the XSDs are used both to describe the contract of the service and to validate the XML serialization of the objects which it uses. Figure 2 details the various services and artifacts related to the description of and process for the transfer of data objects between client and service.

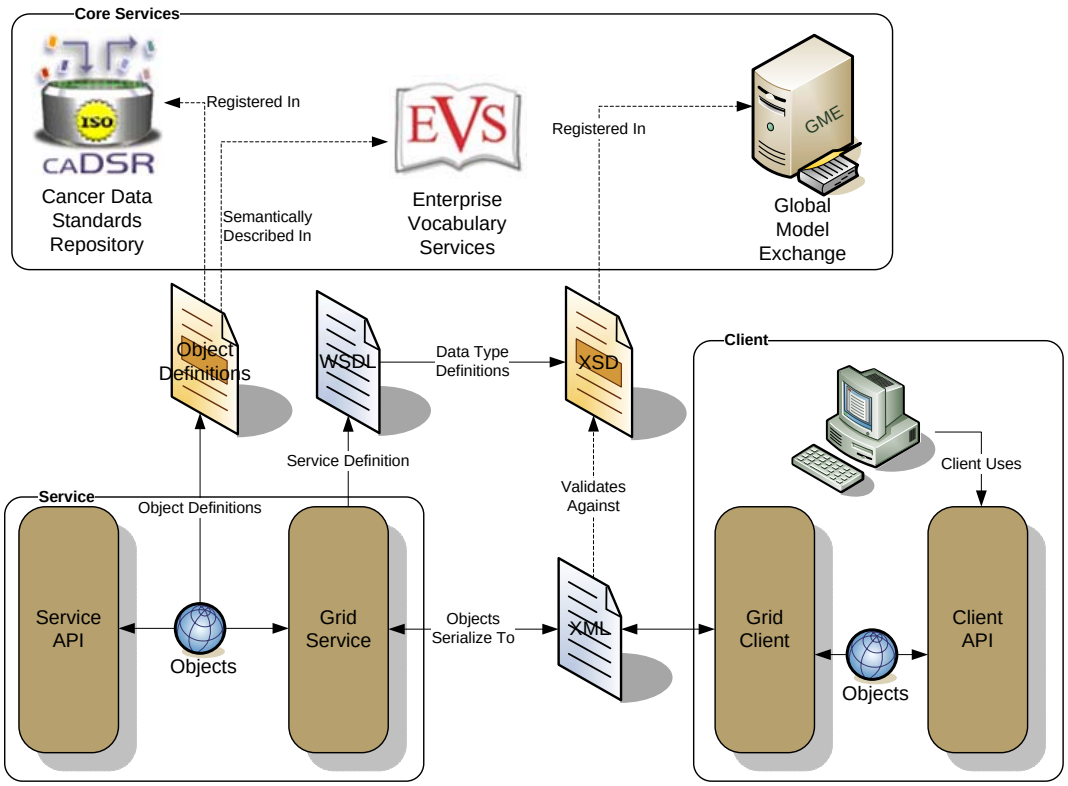

Figure 2 Data Description Infrastructure

Semantically Discoverable

As caBIG aims to connect data and tools from 50+ disparate cancer centers, a critical requirement of its infrastructure is that it supports the ability of researchers to discover these resources. caGrid enables this ability by taking advantage of the rich structural and semantic descriptions of data models and services that are available. Each service is required to describe itself using caGrid standard service metadata. When a grid service is connected to the caBIG grid, it registers its availability and service metadata with a central indexing registry service (Index Service). This service can be thought of as the "yellow pages" and "white pages" of caBIG. A researcher can then discover services of interest by looking them up in this registry. caGrid 0.5 provides a series of high-level APIs and user applications for performing this lookup which greatly facilitate the process.

As the Index Service contains the service metadata (or service data) of all the currently advertised and available services in caBIG, the expressivity of service discovery scenarios is limited only by the expressivity of the service metadata. For this reason, caGrid provides standards for service metadata to which all services must adhere. At the base is the Common Service Metadata standard that every service in caBIG is required to provide. This metadata contains information about the service-providing cancer center, such as the point of contact and the institution's name. Extending beyond this generic metadata are two standards that are specialized for the two types of community-provided services: Data Services and Analytical Services. Both of these standards leverage the data models registered in caDSR and link them to the underlying semantic concepts registered in EVS. The Data Service Metadata details the domain model from which the Objects being exposed by the service are drawn. Additionally, the definitions of the Objects themselves are described in terms of their underlying concepts, attributes, and associations to other Objects being exposed. Similarly, the Analytical Service Metadata details the Objects using the same format as the Data Service Metadata. In addition to detailing the Objects definitions, the Analytical Service Metadata defines the operations or methods the service provides. The input parameters and output of the operations are defined by referencing the appropriate Object definition. In this way, both the data and analytical services fully define the domain objects they expose by referencing the data model registered in caDSR, and identify their underlying semantic concepts by referencing the information in EVS.

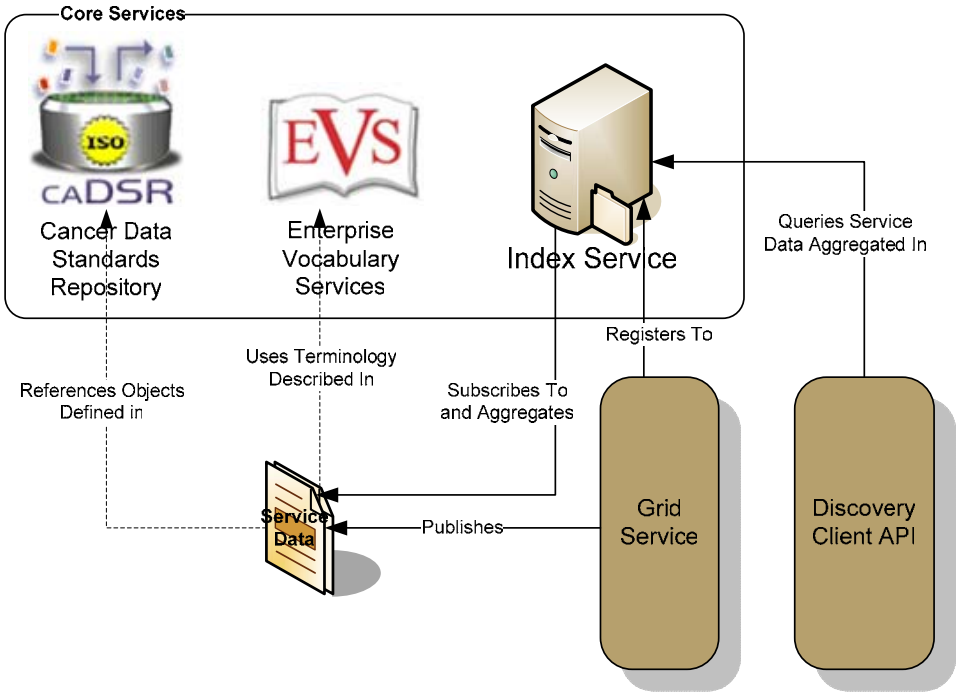

Figure 3 Discovery Overview

As shown in Figure 3, the caGrid discovery API and tools allow researchers to query the Index Service for services satisfying a query over the service metadata. That is, researchers can lookup services in the registry using any of the information used to describe the services. For instance, all services from a given cancer center can be located, data services exposing a certain domain model or objects based on a given semantic concept can be discovered, as can analytical services that provide operations that take a given concept as input.

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 systems for authenticating and authorizing requests to the corresponding data sources.

The caGrid security architecture, shown below in Figure 4, is compromised of several components. Components can be classified as *core components* and *external components*. Core components are required by the architecture and are essential for meeting the security requirements for caBIG. External components are those that are considered extensions of the

core security architecture. The caGrid security architecture is composed of five core components:

1. **Grid Security Infrastructure (GSI):** Globus provided security infrastructure.
2. **Authorization Manager:** Authorization callback mechanism.
3. **Grid User Management Service (GUMS):** Grid Service for the management and creation of grid users and grid user credentials.
4. ***Grid Virtual Organization Service (GVOS)*** *[not implemented in caGrid 0.5]*: Grid Service for the management of virtual organizations.
5. **caGrid Attribute Management Service (CAMS):** Grid Service for the management of user/virtual organization attributes.

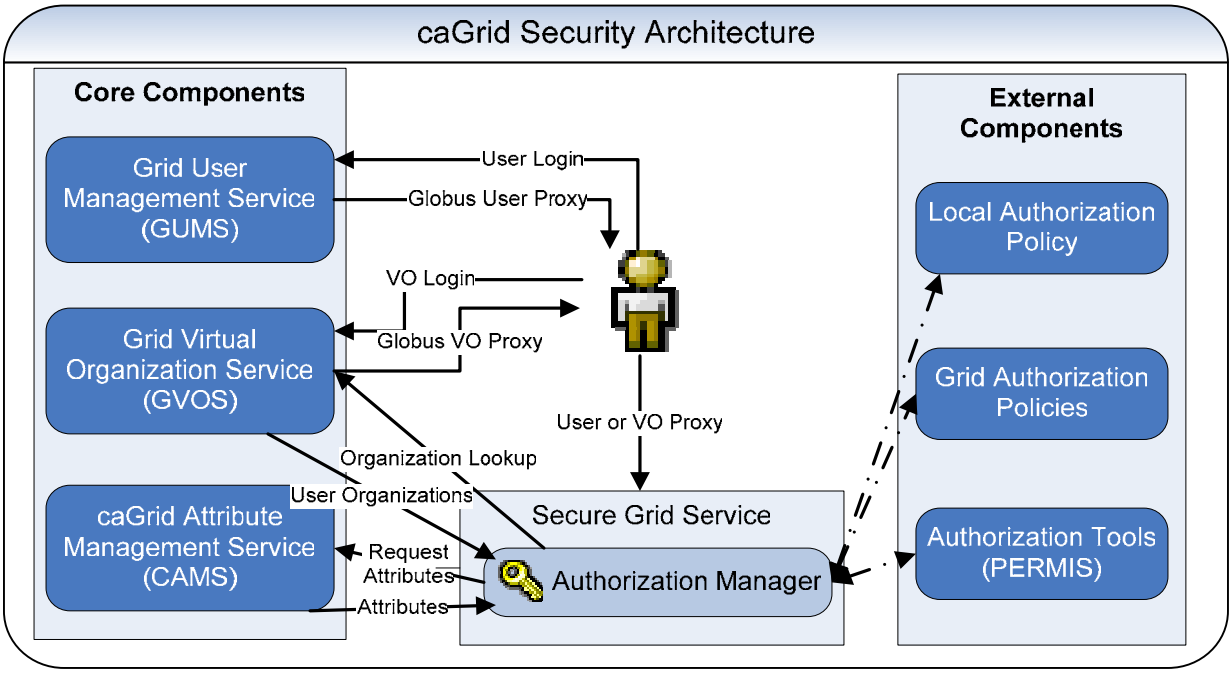

Figure 4 caGrid Security Architecture

Service Nodes (Reference Implementation)

The caGrid 0.5 release also provides some examples of existing community provided service nodes for the test bed. The services nodes includes ***PIR***, ***rProteomics, caArray, caBIO and caTIES.***

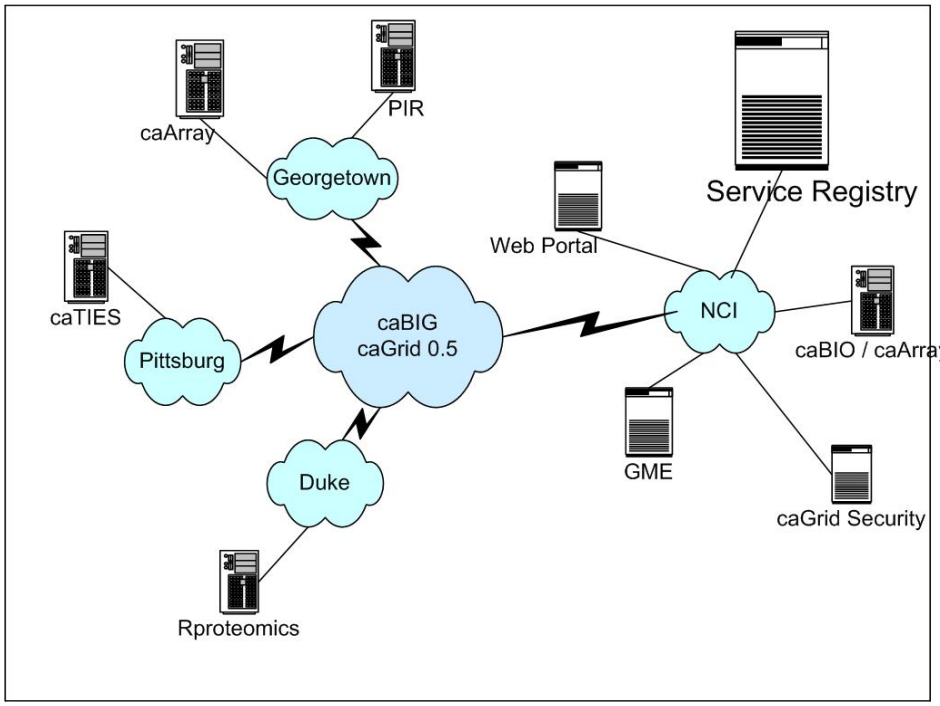

Figure 5 caGrid 0.5 Test Bed Infrastructure (Including the Service Nodes)

Service Node # 1: PIR (Data Service)

Provider of PIR – Data Service: Georgetown University

About PIR: Protein Information Resource (PIR) is an Integrated Protein Informatics Resource for Genomic/Proteomic Research. It consists of three parts. UniProt Universal Protein Resource is the Central Resource of Protein Sequence and Function. PIRSF Family Classification System is the Protein Classification, and Functional Annotation and iProClass Integrated Protein Knowledgebase is the Data Integration and Functional Analysis.

Service Node # 2: caArray (Data Service) – Instance # 1

Provider of caArray – Data Service: Georgetown University

About caArray: caArray is a standards based data repository of microarray experiment data using MIAME standard. The MIAME standard describes the **M**inimum **I**nformation **A**bout a **M**icroarray **E**xperiment that is needed to enable the interpretation of the experiment results unambiguously and, potentially, to reproduce the experiment.

Service Node # 3: caArray (Data Service) – Instance # 2

Provider of caArray – Data Service: ***NCICB***

About caArray: caArray is a standards based data repository of microarray experiment data using MIAME standard. The MIAME standard describes the **M**inimum **I**nformation **A**bout a **M**icroarray **E**xperiment that is needed to enable the interpretation of the experiment results unambiguously and, potentially, to reproduce the experiment.

Service Node # 4: rProtemics (Analytical Service)

Provider of rProteomics – Analytical Service: ***Duke University***

About rProteomics: rProteomics is to find biomarker and to build predictive model. It develops analytical routines for proteomics data like denoising, background removal, peak identification, spectral alignment, normalization, peptide quantitation. Its focus is on analytics rather than databases, LIMS, protein identification. RProteomics is a critical step in the proteomics pipeline LIMS -> repository -> RProteomics -> classification -> protein identification, and RProteomics provides integration of Q5 classification.

Service Node # 5: caTIES (Data Service)

Provider of caTIES – Data Service: ***University of Pittsburg – Medical Center***

- ***About caTIESs:*** The Cancer Text Information Extraction System (caTIES) caTIES is a text processing system that creates de-identified structured data from unstructured free-text pathology reports and makes reports accessible to researchers.
  - Information about tumor, stage, prognostic factors
  - Index to fixed tissue, source of annotation for frozen or processed tissue

  caTIES de-identifies entire corpus of reports, creates concept codes using NCI metathesaurus, to MySQL datastore. Deployed to adopter at University of Pennsylvania, intention is to create a network of institutions that can share data and tissue

Service Node # 6: caBIO (Data Service)

Provider of caBIO – Data Service: ***NCICB***

About caBIO: caBIO provides standard object models and a uniform programmatic interface access (Java, web services, Perl) to the entire caCORE technologies. It also provides an abstraction layer that allows developers to access genomic, systems biology, clinical and pre-clinical, biomedical metadata and a wide variety of medical vocabularies.

Acknowledgements

Thanks to the following who made caGrid 0.5 possible

- **caGrid team**
  - **Ohio State University - Biomedical Informatics Department**
    (Arthur G. James Cancer Hospital & Richard J. Solove Research Institute)
    - Scott Oster - Shannon Hastings -Steve Langella
    - Tahsin Kurc - Joel Saltz
  - **Panther Informatics (NCICB Contractor)**
    - Nick Encina
  - **SAIC (NCICB Contractor)**
    - William Sanchez - Tara Akhavan - Manav Kher
    - Rouwei Wu - Jijin Yan
- **NCICB**
  - Peter Covitz - Avinash Shanbhag - George Komatsoulis
  - Denise Warzel - Frank Hartel
- **TerpSys (NCICB Contractor)**
  - Gavin Brennan - Troy Smith - Wei Lu - Doug Kanoza
- **Booz Allen Hamilton (caBIG Project Team)**
  - Arumani Manisundaram - Mike Keller - Brian Davis

Thanks to the following who made reference implementations possible.

- ***Georgetown University – PIR Project***
  - Baris Suzek - Scott Chung - Cathy Wu - Peter McGarvey
  - Hongzhan Huang - Sehee Chung - Hsing-Kuo Hua
- ***Georgetown University - caArray Project***
  - Colin Freas - Nick Marcou - Arnie Miles
  - Jess Cannata - Jack Yuelin Zhu
- ***Duke - rProteomics Project***
  - Salvatore Mungal - Patrick McConnell
- ***University of Pittsburg – Medical Center - caTIES Project***
  - Rebecca Crawley - Kevin Mitchell - Linda Schmandt
  - Aditya P Nemlekar - Girish Chavan
- ***NCICB / SAIC - caArray Project***
  - John Moy - Sumeet Muju - Juergen Lorenz
- ***NCICB / SAIC***
  - Andrew Shinohara - Mike Connelly - Jennifer Zeng
  - Nafis Zebarjadi