caBIG®

cancer Biomedical
Informatics Grid®

Publication and Consumption of caBIG Data Services using .NET

Marty Humphrey
University of Virginia

Microsoft eScience Workshop 2008

December, 2008

caBIG™ Vision and Goals

caBIG™ Vision

A virtual network of interconnected data, individuals, and organizations that whose goal is to redefine how research is conducted, care is provided, and patients/participants interact with the biomedical research enterprise.

caBIG™ Goals

- **Adapt or Build** tools for collecting, analyzing, integrating and disseminating information associated with cancer research and care
- **Connect** the cancer research community through a shareable, interoperable electronic infrastructure
- **Deploy and Extend** standard rules and a common language to more easily share information

From Ken Buetow, NCI

caBIG™ Core Principles

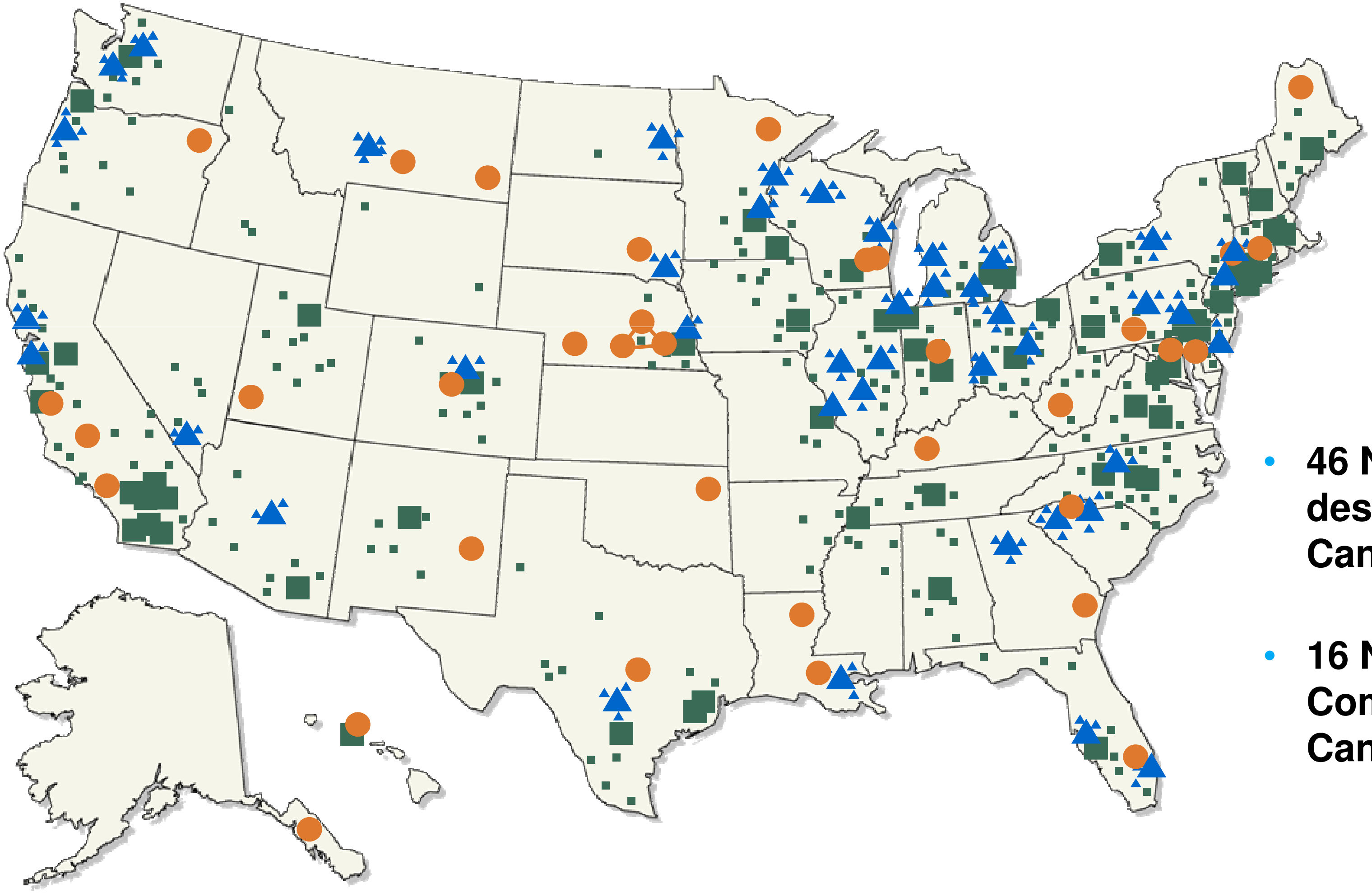
- **Open Access** – caBIG™ is open to all, enabling wide-spread access to tools, data, and infrastructure
- **Open Development** – Planning, testing, validation, and deployment of caBIG™ tools and infrastructure are open to the entire research community
- **Open Source** – The underlying software code of caBIG™ tools is available for use and modification
- **Federation** – Resources can be controlled locally, or integrated across multiple sites

From Ken Buetow, NCI

caBIG™ Deployment:

Adoption is Well Underway Nationally

NCI-Designated Cancer Centers, Community Cancer Centers, and Community Oncology Programs



- **46 NCI-designated Cancer Centers**
- **16 NCI Community Cancer Centers**

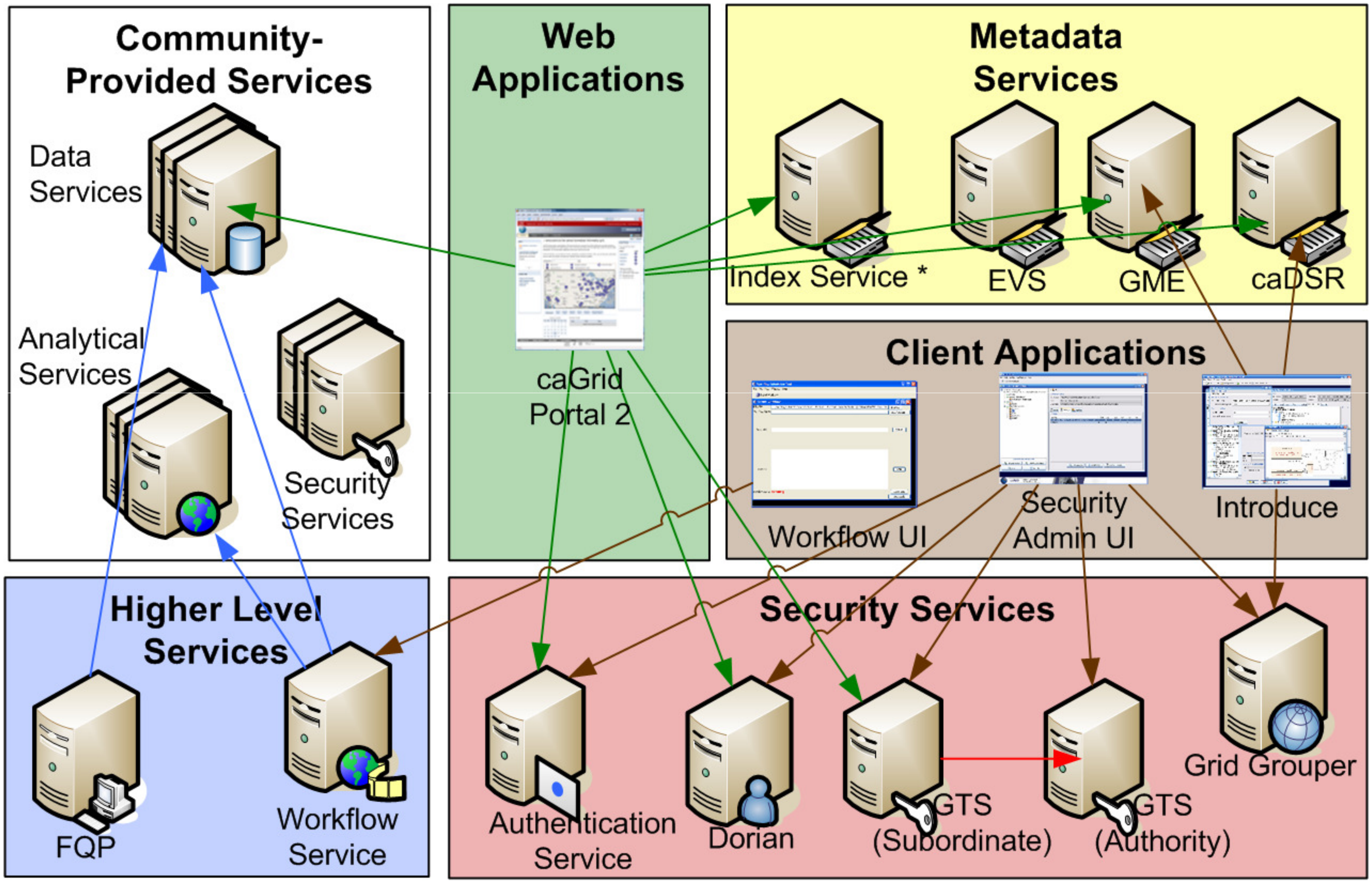
From Ken Buetow, NCI

What is caGrid?

- A grid based **software infrastructure** consisting of services, toolkits, APIs, and applications
- A **production grid deployment** of the core services provided by that infrastructure
- A **community of developers** leveraging that grid and infrastructure to provide applications and services to the cancer research community

From Scott Oster, Ohio State University

caGrid Production Environment

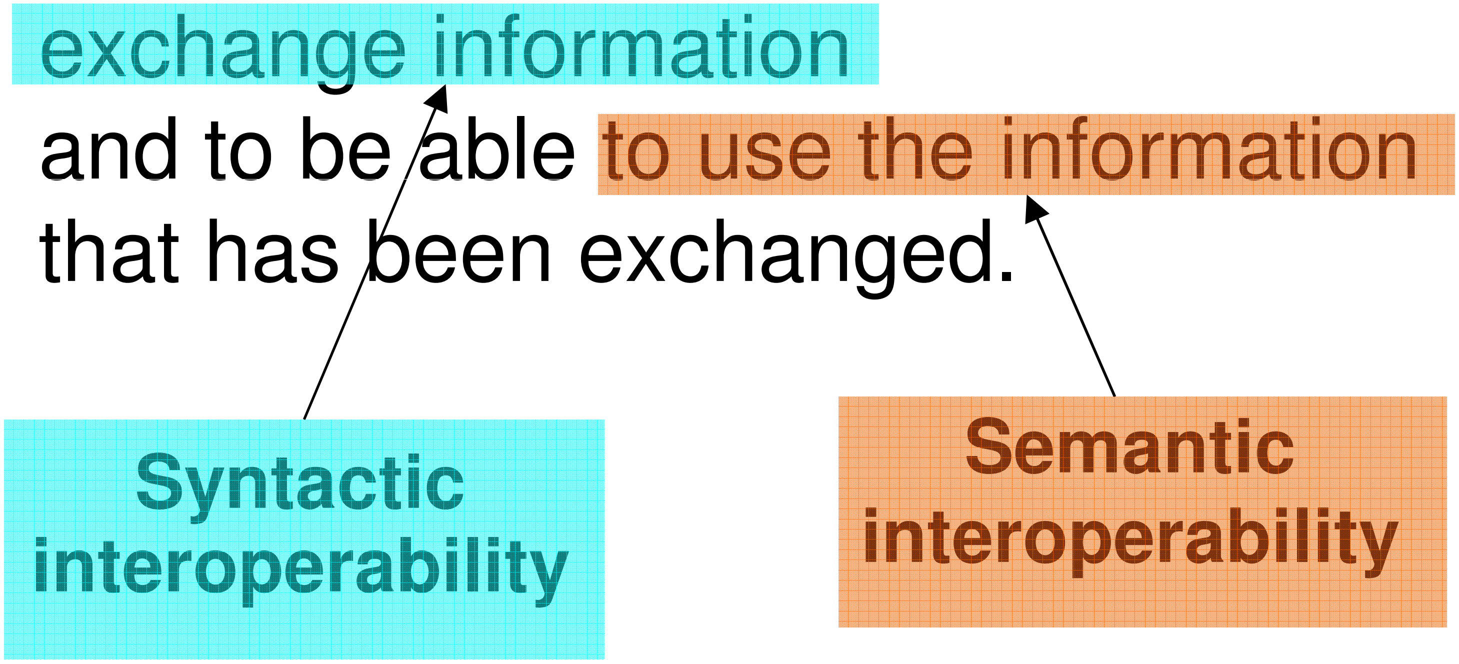


*All Services Register with the Index Service

From Scott Oster, Ohio State University

Interoperability

- The ability of multiple systems to exchange information and to be able to use the information that has been exchanged.



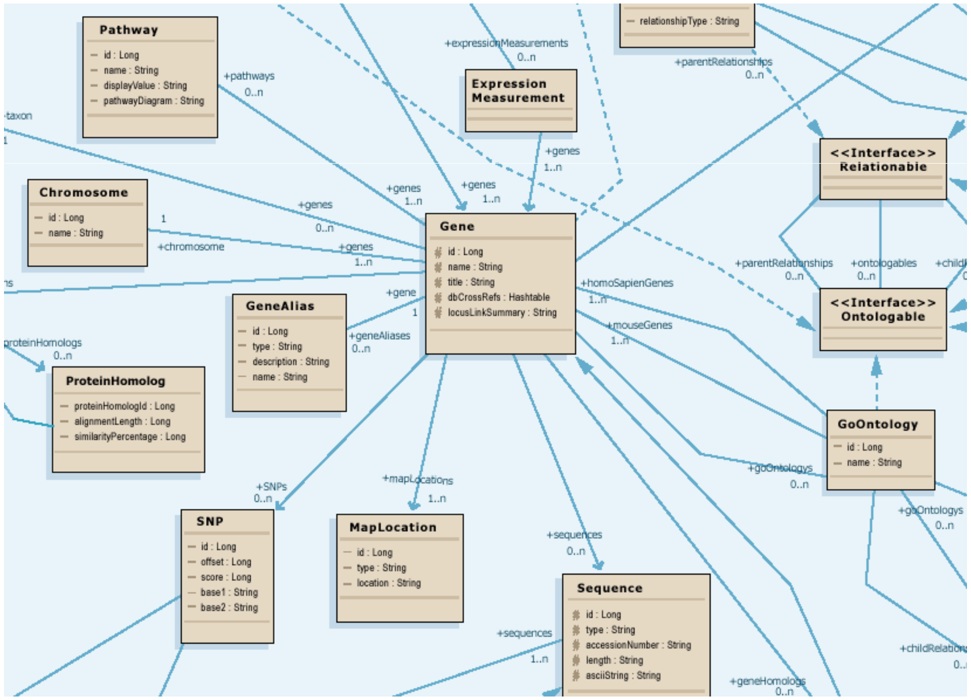
Syntactic interoperability

Semantic interoperability

From Scott Oster, Ohio State University

Modeling for Interoperability

- **Class diagram models target domain**
- **Logical model is basis for semantic integration**
- Focus on attributes and relationships of domain objects



From Scott Oster, Ohio State University

Data Model Meaning

- 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

From Scott Oster, Ohio State University

Metadata Services

- **Cancer Data Standards Repository (caDSR)**
  - caBIG projects register their data models as Common Data Elements (CDEs) which are semantically harmonized and then centrally stored and managed the caDSR
  - The caDSR grid service provides:
    - Model discovery and traversal
    - caGrid standard metadata generation capabilities
- **Enterprise Vocabulary Services (EVS)**
  - EVS is set of services and resources that address the need for controlled vocabulary
  - The EVS grid service provides:
    - Query access to the data semantics and controlled vocabulary managed by the EVS
- **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.
  - The GME grid service provides:
    - Access to the authoritative structural representation of data types on the grid
- **Globus Information Services: Index Service**
  - The Globus Information Services infrastructure provides a generic framework for aggregation of service metadata, a registry of running Grid services, and a dynamic data-generating and indexing node, suitable for use in a hierarchy or federation of services
  - The Index grid service provides:
    - Yellow and white pages for the grid

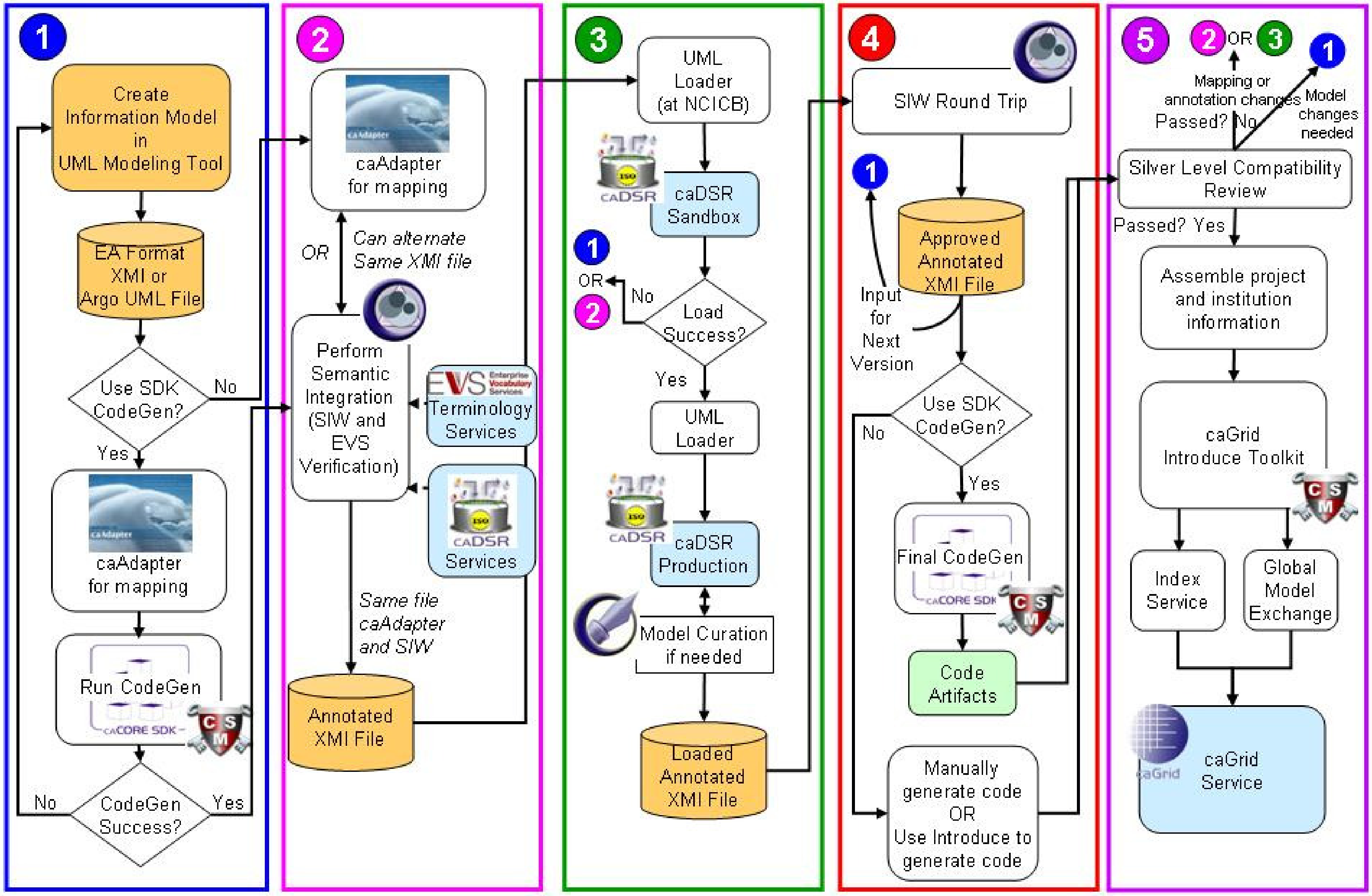
From Scott Oster, Ohio State University

Why .NET?

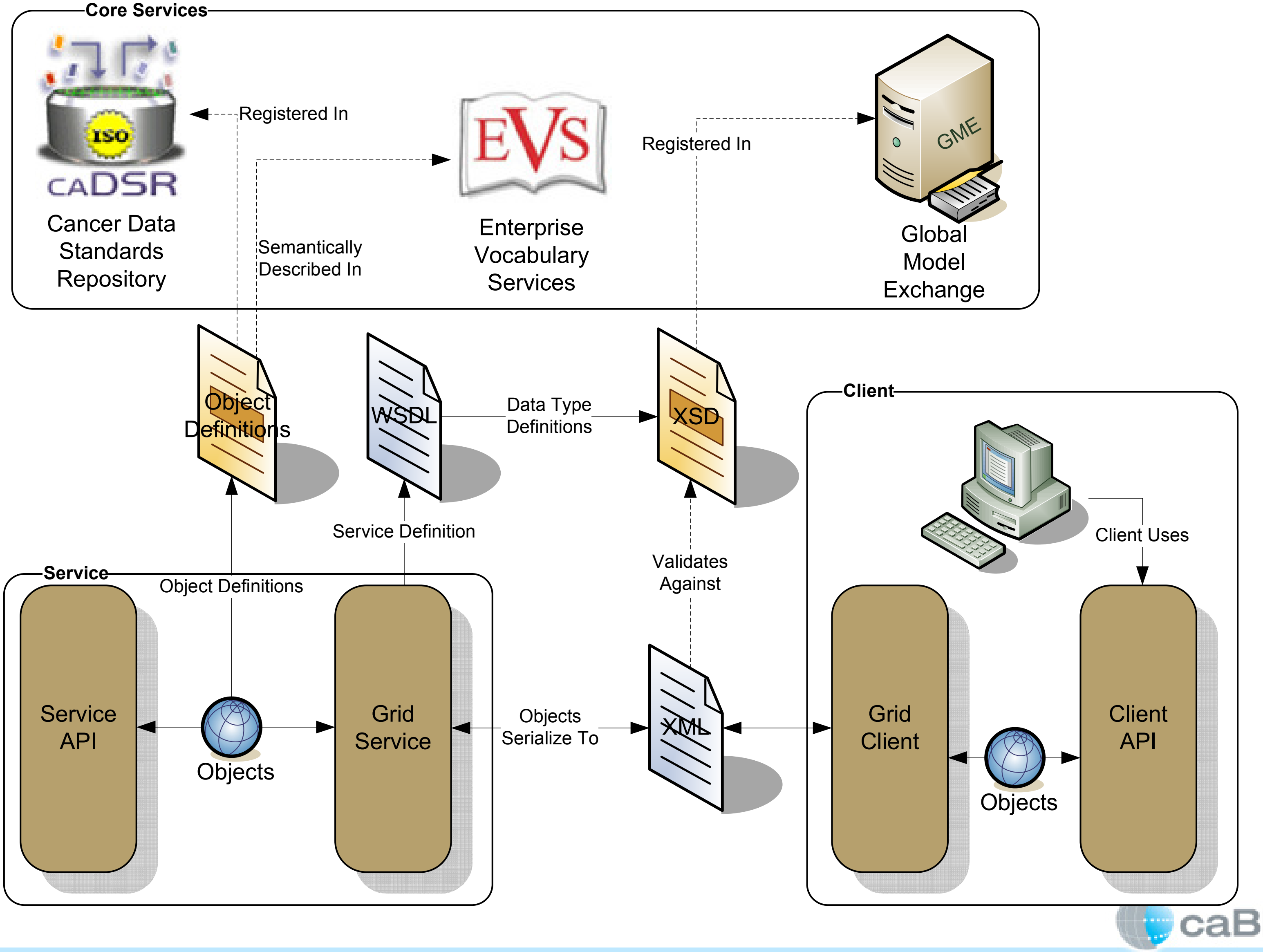
- **Give existing .NET-based developers/infrastructure easy way to participate in caBIG**
- **Give new developers a CHOICE!**

- **Leverage .NET/Windows ecosystem today:**
  - Visual Studio, .NET, SQL Server, Windows Workflow Foundation, LINQ
- **Leverage .NET/Windows ecosystem in the future:**
  - Sharepoint, Hyper-V, Cloud computing, Microsoft Parallel Computing Initiative, Modeling: Project OSLO

caCORE SDK centric caGrid data service development



caBIG Clients and Services



A Scientific User Scenario

- A researcher is studying human BRCA1 gene and wants to find information available in public resources on protein encoded by this gene

caBIG™ Translation of the User Scenario

1. **Discover multiple caGrid Data Services providing Protein information**
   - Use caGrid Discovery Client
2. **Find how to combine the information from these Data Services**
   - Find semantically equivalent data elements (Common Data Elements) from different data services
3. **Identify/query the Protein corresponding to BRCA1 gene**
   - Run caBIG™ Query Language (CQL) queries using caGrid Data Service Client
4. **Collect information on the <u>same</u> protein from different resources**
   - Run multiple or federated CQL queries against different Data Services leveraging Common Data Elements

DEMO: Building a .NET Client for a caBIG Data Service

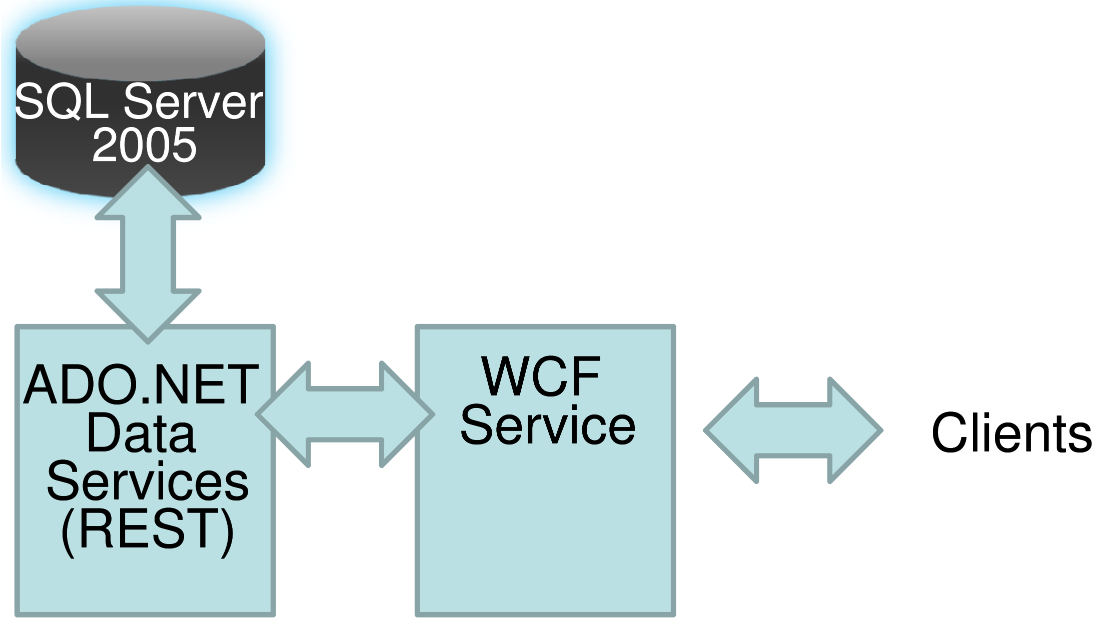
Demo Recap (1/2)

1. **Generate proxies from service**
   1. Get all WSDL and XSD from tool: SvcUtil.exe
   2. Modify WSDL in 6 places (QueryResourceProperties, GetMultipleResourceProperties, GetResourceProperty)
   3. Generate proxy code via SvcUtil.exe
2. **In VS**
   1. Add CaBIOSvc.cs and output.config (as app.config)
   2. Add references: System.ServiceModel and System.Runtime.Serialization
   3. Add code
3. **Run**

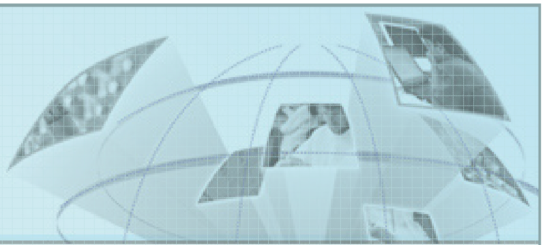
Demo Recap (2/2): Results

```
<ns2:Gene id="9188" fullName="Breast cancer 1, early onset"
   clusterId="194143" symbol="BRCA1"
   xmlns:ns2="gme://caCORE.caCORE/3.1/gov.nih.nci.cabio.domain" />
<ns3:Gene id="137079" fullName="Breast cancer 1" clusterId="244975"
   symbol="Brca1"
   xmlns:ns3="gme://caCORE.caCORE/3.1/gov.nih.nci.cabio.domain" />
<ns4:Gene id="1685" fullName="Breast cancer 2, early onset"
   clusterId="34012" symbol="BRCA2"
   xmlns:ns4="gme://caCORE.caCORE/3.1/gov.nih.nci.cabio.domain" />
<ns5:Gene id="136510" fullName="Breast cancer 2" clusterId="236256"
   symbol="Brca2"
   xmlns:ns5="gme://caCORE.caCORE/3.1/gov.nih.nci.cabio.domain" />
```

.NET caBIO Data Service



1. CQL
2. WSRF
3. caBIO
4. Index Service



DEMO: Building a .NET Service for caBIO Data

Demo Recap (1/2)

- **Get data into SQL Server:**
  - Easy, once we figured out how to do it 🙂
- **Conform to Data Service WSDL**
  - Proxy-gen after WSDL mods (6 lines)
- **Get data out of SQL Server**
  - ADO.NET Data Services: REST service **(<u>nice</u>)**
- **Write CQL processor**
  - A challenge so far… only minimal functionality implemented right now

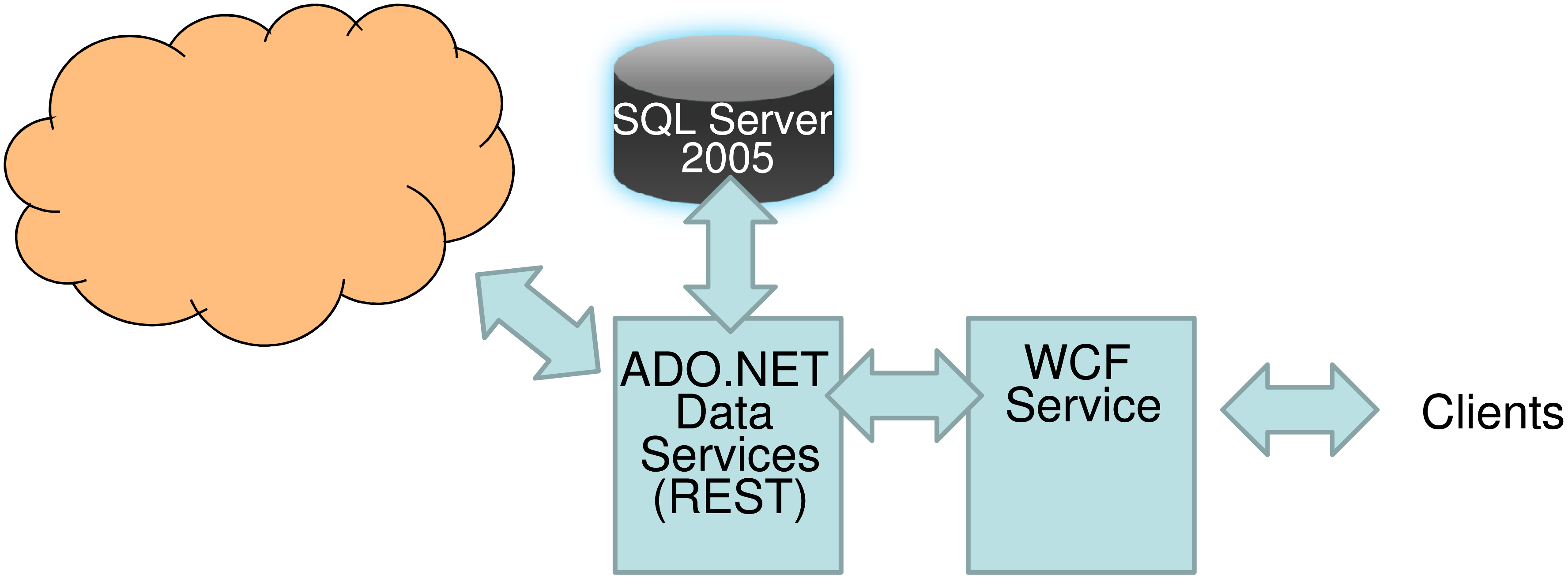
Demo Recap (2/2)

- **Implement WSRF methods**
  - Surprisingly, so far so good
- **Interact with GME**
  - Challenging: WSDL is not as “expressive” as other services
  - Must reverse-engineer the protocol (a continuing issue)
    - Looking for the new version of GME…
- **Publish to Index Service**
  - Okay, but not complete (GetResourceProperty: DomainModel and ServiceMetadata)

- **Aim demo client at new service**

DEMO: Accessing a Deployed .NET Service for caBIO Data using the caGrid Portal

.NET caBIO Data Service



1.CQL
2. WSRF
3. caBIO
4. Index Service

VERY Preliminary Performance*

	Local (SQL Server 2005)	Cloud (Azure: SQL Data Services)
“How many CHROMOSOMES?” (84)	1 second	1 second
“How many GENES?” (202250)	68 seconds (LINQ) (“count” is not supported in LINQ → ADO.NET Data Services)	198 seconds (in 405 chunks of 500)
“How many GENES?” (max: 500)	1 second (LINQ) 19 seconds (REST)	2 seconds
“Find me the GENEs like BRCA” (4)	2 seconds	3 seconds

.NET-based Services: Status

- **Tutorial has just been completed**
- **Continuing issues:**
  - CQL processor
  - Interacting with GME / caDSR
- **Future work:**
  - Consider Analytical Services
  - Security

Summary

- .NET ecosystem has significant potential to caBIG participants
- .NET Working Group has begun a sustained effort at extending/leveraging this .NET ecosystem
- Strong early successes with clients and caBIO Data Service
- Much more work necessary to move beyond prototyping phase
  - Improve ease-of-use
  - Integrate with caGrid security infrastructure
  - Provide support for early adopters