National Institutes of Health
National Center for Research Resources

caBIG™ Overview

May 2006

The NIH National Center for Research Resources has contracted the MITRE Corporation to track developments and to inform the research community in the area of clinical research information technology through a series of targeted research reports.

MITRE
Center for Enterprise Modernization
McLean, Virginia

NIH National Center for Research Resources (NCRR)

NCRR provides laboratory scientists and clinical researchers with the environments and tools they need to understand, detect, treat, and prevent a wide range of diseases. With this support, scientists make biomedical discoveries, translate these findings to animal-based studies, and then apply them to patient-oriented research. Ultimately, these advances result in cures and treatments for both common and rare diseases. NCRR also connects researchers with one another, and with patients and communities across the nation. These connections bring together innovative research teams and the power of shared resources, multiplying the opportunities to improve human health. For more information, visit www.ncrr.nih.gov.

Accelerating and Enhancing Research From Basic Discovery to Improved Patient Care

The MITRE Corporation

The MITRE Corporation is a private, independent, not-for-profit organization, chartered to work solely in the public interest. MITRE manages three federally funded research and development centers (FFRDCs) and partners with government sponsors to support their critical operational missions and address issues of national importance. For more information about The MITRE Corporation and its work, visit www.MITRE.org.

Applying Systems Engineering and Advanced Technology to Critical National Problems

Table of Contents

List of Figures

List of Tables

1. Introduction

Throughout the academic medical community and the healthcare industry there is a clear imperative to improve the connection between basic research and patient care. In 2004, the National Cancer Institute (NCI) National Center for Bioinformatics (NCICB) initiated a pilot project to develop an information infrastructure that enables this connection. The Cancer Biomedical Informatics Grid (caBIG) seeks to connect the entire cancer community, from bench scientists to cancer clinicians to the Food and Drug Administration (FDA). In the brief time since this pilot project began, caBIG has gained awareness and interest from both within and beyond the cancer community. This paper describes caBIG, discusses its current strategy, and provides the non-cancer research community with a perspective on the potential applicability of caBIG to their particular areas of interest.

2. caBIG Vision

"...if you don't have a framework for a vision, nobody moves in a really new direction. Today there are genuinely new opportunities, but the new science will require interdisciplinary collaborations. That will be key." National Institutes of Health (NIH) director Elias Zerhouni, M.D.[1]

> ***Vision:*** *"caBIG will become a self-sustaining network, which will foster improvements in collaborative projects and increase the speed and efficacy of treatment to benefit patients."*
>
> ***Mission:*** *"caBIG participants will develop readily disseminated standards tools and information systems for the management of clinical and research activities in oncology. These will include systems for the management of cancer clinical trials, standards for integrative research systems, a coherent approach to biospecimen informatics management, and the underlying architectures, vocabularies and data elements that will facilitate sharing and access to these systems."*[2]

The purpose of caBIG is to enable interdisciplinary collaborations across the field of cancer research by providing an informatics infrastructure to accelerate the pace and translation of scientific discovery to prevent and treat disease. This collaboration is increasingly necessary because new tools, such as genomics and proteomics, are revolutionizing our ability to understand the complex interactions within the body that contribute to health and disease. The new tools are changing the content of science, the structure of research teams, the way that information is collected and exchanged, and the processes by which scientific investigations are carried out.[3]

Impacting patient care via basic science research requires contributions from participants who have widely diverse backgrounds, including: clinicians, biochemists, information technology professionals, administrators, financial specialists, pharmacologists, molecular biologists, ethicists, legal professionals, and patients. Traditionally, these have been viewed as distinctly independent disciplines, and the pace of advancing patient care has reflected a slow coalescing of understanding that is played out in a large body of peer-reviewed literature. To date, the published journal article remains the primary vehicle for sharing information between disciplines and across institutions.

Sharing information and data at far earlier stages holds the promise of accelerating the pace of translational medicine by facilitating the discovery of related activities and allowing them to dynamically come together, regardless of the disciplines or institutions involved. This capability, however, would require that electronic datasets and software tools be readily useable across disciplines and institutions. Achieving this level of interoperability is difficult because each discipline has developed independent vocabularies and terminologies, highly tailored information management systems, and customized analysis tools. Even within a single institution, working across these systems can be very difficult; crossing institutional boundaries is generally prohibitive.

caBIG strives to overcome these challenges for cancer research by creating an information infrastructure that enables broad interoperability, both within and across institutional boundaries.

It seeks to establish interoperability at multiple levels—from machine-machine exchanges of bits across a network, to the compatibility of tools, to accurate exchanges of complex concepts—enabled through the use of standardized definitions and semantics. And there are other potential benefits, aside from the exchange of scientific information. For example, every grantee must provide

reports to the National Institutes of Health (NIH) annually; generating these often involves custom programming that incurs substantial costs. If caBIG's datasets and tools were able to generate these reports automatically, the savings could be substantial, resulting in increased resources available for research.

caBIG's strategic infrastructure involves developing open source software and using commercial software where appropriate to provide interoperability, connecting users and tools through a grid infrastructure, and providing researchers with a set of grid-enabled applications. The software tools support basic, clinical and translational research functions, such as sharing tissue bank data. Providing the capability for researchers to use caBIG-enabled applications will help remove technological barriers to sharing data on the grid. Another key part of the strategy is using well-defined standards for data exchange and developing a core vocabulary so that results can be compared. caBIG will be used by NIH, cancer researchers worldwide, and other biomedical researchers who are not part of the cancer community. caBIG tools and infrastructure components are freely available on the caBIG Web site [https://caBIG.nci.nih.gov].

3. Community and Organization

caBIG is in its third year of active development. One of the first, and very crucial, tasks has been to develop a community of researchers and informaticians from government, academia, and industry to develop the strategy and infrastructure.

Initially, a core group of Cancer Centers at academic medical centers were contracted to develop the first set of caBIG components[4]. Users can also contribute tools and data to the caBIG project. There are currently over 800 caBIG participants, drawn from Cancer Centers, academic medical centers, clinical research networks, government, and industry. It is important to recognize that many of the participants are unpaid volunteers who are contributing their time and effort to caBIG to advance the achievement of the vision.

In order to develop caBIG, a stakeholder strategy team was formed. The team visited over 50 NCI-designated Cancer Centers to gather their priorities for research tools and other IT resources, as shown in the figure below. As a result of this analysis, community stakeholders formed groups to begin building out the key tools that were perceived as being most needed to advance research and to provide the critical mass of technology that would encourage widespread adoption of caBIG.

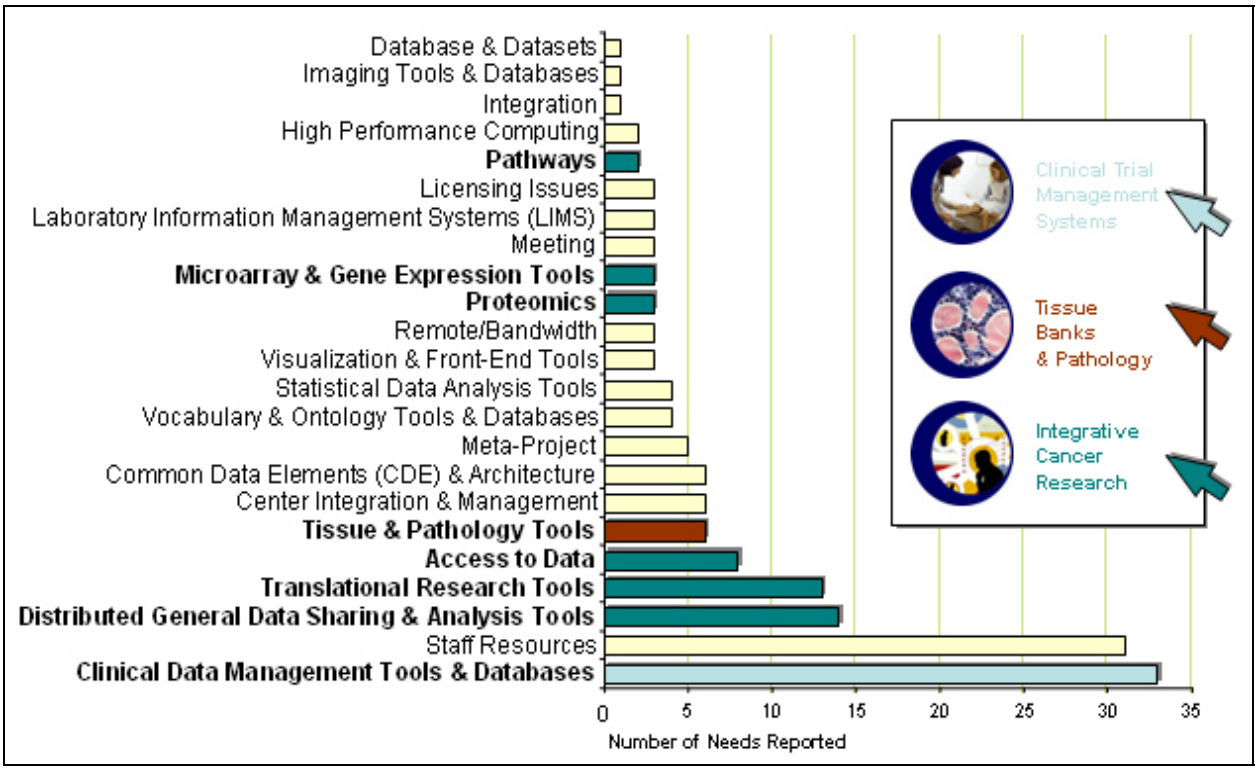

Figure 3-1. caBIG Community Needs[1]

[1] Figure provided by caBIG

The caBIG community is organized into nine workspaces.[5] These workspaces have teams of contractors and volunteer subject matter experts who prioritize objectives and shape the requirements of the developed infrastructure. The workspaces include:

- Strategic-level workspaces:
  - **Strategic Planning:** comprised of members selected from all other workspaces, it sets the strategic direction for the community
  - **Data Sharing and Intellectual Capital:** is identifying legal, regulatory, and intellectual property factors that impact data sharing and resource sharing
  - **Training:** is developing training and documentation materials
- Crosscutting workspaces:
  - **Architecture:** is developing the architectural standards and Grid infrastructure, including establishing and enforcing security policies
  - **Vocabularies and Common Data Elements (VCDE):** is responsible for overseeing vocabulary, ontology, and common data elements (CDE), and for developing standards for the representation of these core elements
- Domain workspaces:
  - **Integrative Cancer Research (ICR):** is working on tools to store, manage, and analyze cell and molecular data generated by high-throughput genomic and proteomic techniques, and correlating the results with information contained in the other workspaces
  - ***In Vivo* Imaging:** is creating software tools and modeling methods to manage and analyze imaging data
  - **Clinical Trial Management Systems (CTMS):** is developing a comprehensive tool suite for clinical trial management in the cancer research community
  - **Tissue Banks and Pathology Tools (TBPT):** is developing tools that enable specimen and information sharing across cancer centers

Special interest groups (SIG) are formed from subsets of the participants within each of the workspaces. To help identify where tighter coordination among groups is needed, liaisons are chosen to work across SIGs within a workspace, and also between the workspaces. Some workspace representatives are under contract with NCI to ensure stabilization and participation in the workspaces.

Within the workspaces, community needs are discussed and prioritized in order to guide new software development. Each software development project has a developer creating the tool, and one or more formal adopters providing feedback on its use. Both the developers and formal adopters are under contract to NCI.

4. Basic and Translational Research Support

Consider the following scenario:

> A prostate cancer researcher seeks to elucidate the molecular basis of a particular phenotype in a particular patient population: tumor vascularization in Asian males age 40-45. First, suppose she is able to augment the relatively meager amount of tumor and matched-control tissue available at her home Cancer Center with material available throughout the entire network of 61 Cancer Centers. A simple interface lets her search pathology reports to identify and request relevant specimens, automatically filtering out those that may lack the appropriate subject consent and Institutional Review Board (IRB) approval. As all samples are de-identified by the institution responsible for maintaining them, both HIPAA and IRB issues are expedited and are not a barrier to the inquiry. Originating labs receive the request electronically and contact the researcher to discuss the study and attribution. With parameters of the collaboration worked out, a software tool quickly identifies the location of the requested sample, and it's shipped out.
>
> Next, suppose the researcher processes this tissue with high-throughput genomic and proteomic techniques to produce electronic datasets. Unfortunately, the investigator discovers that the number of matched-control samples is below that needed for the required statistical power. A query across the grid is used to find genomic and proteomic data that has been previously collected and can be re-purposed for use as matched-control data in this study; again, the parameters of collaboration are worked out with the data originators. Importantly, the manner in which the legacy genomic and proteomic data has been represented and stored affords seamless integration with the newly generated data.
>
> Using novel computational algorithms available at a sister cancer center, the data is analyzed remotely. Transcriptional networks and biochemical pathways involved in vascularization are mapped out. A particular allele with high prevalence in Asian males is identified and found to produce an end-product protein that is required for tumor vascularization. Follow-up research reveals that this result is specific to Asian males having this particular allele.
>
> A publication that includes proper attribution of all participants is submitted to a high-profile journal. All collaborators benefit from the acknowledgments and attributions published with the article. In addition, annotation of this allele with its phenotypic behavior is now available through a form, and this information is disseminated much faster via the electronic infrastructure than through the release of the publication. Before the hard copy of the publication arrives, other researchers pick up the challenge of identifying small molecules that can inhibit protein activity, and an investigator-initiated clinical trial application is submitted to the FDA.

Today, a prostate cancer researcher would spend months tracking down a sufficient supply of tissue tumor samples. Genomic and proteomic datasets would be useable only by the lab of origin, and analysis tools would be limited to those available locally. Research results would find their way to interested parties only after a lengthy process of publication and conference presentations. caBIG infrastructure and tools for basic and translational research seeks to change all this, accelerating the translational research enterprise by allowing investigators to have unprecedented access to life sciences data and information much more quickly than ever before.

Table 4-1. Basic and Translational Research Tasks & Benefits

Representative Tasks	Without caBIG	With caBIG	Benefit
1. Management of reagents and specimens	• Done with Excel spreadsheets • Done with laboratory information management systems (LIMS), which cannot exchange data with each other or link with clinical systems	• caLIMS manages lab workflow • caTISSUE (the cancer tissue database) manages the entire tissue banking system	• Storage and processing of lab and specimen data and metadata becomes standards-based, enabling interoperability across labs and institutions • Lab data can be connected to de-identified clinical data longitudinally
2. Management of molecular and cellular and imaging data	• Primarily done with customized systems that generally fail to interoperate • Emerging standards struggle for broad adoption without strong tool support	• Range of information management systems and developed tools that drive standards adoption: – caArray used to store and manage DNA microarray data in a manner that is compliant with Minimum Information About a Microarray Experiment (MIAME) and Microarray and Gene Expression Markup Language (MAGE-ML) – Proteomics LIMS used to manage proteomics	• Representations of genomics, proteomics, and imaging data and metadata become standards-based, supporting the aggregation of this data across institutes: – Team science is truly enabled. – Aggregated data improves statistical power, driving new scientific discovery – More value is returned on the investment made to collect experimental data
3. Placing data in a broader biological context	• Non-interoperable tools that have no underlying standards are used to: – Author gene and protein sequence, pathway, and image annotations – Query annotations – Place molecular data in a pathway context – Construct or infer pathway elements • The lack of standards and interoperability results makes it difficult to query in an automated fashion	• Standardized annotation formats and vocabulary are used for gene and protein sequences, pathways, and imaging data • Developed tools support these standard in authoring and querying of annotations: – Function Express for probe annotation on microarrays – Pathway Tools for annotation of pathway data – Reference Research Workbench for imaging data	• Annotation data can be queried in a much more sophisticated manner, getting the investigator the right information quickly • Data are formatted and compatible across tools and reported results—no additional programming is required

Representative Tasks	Without caBIG	With caBIG	Benefit
4. Biomarker discovery	• Data from cell and molecular experiments are linked with the corresponding clinical data in an ad hoc manner: – Labor-intensive, hand-linking of the data to match clinical data with the bench-science • Computational tools for discovering statistically meaningful correlations between molecular events and patient health can be difficult to find and use	• Interoperating caBIG tools allow within-subject clinical and bench-science experimental data to be linked: – Standardized vocabularies are used across data types – Subject privacy is maintained in a rigorous, well-vetted manner – Data is structured in a standardized manner • Computational tools are widely available	• Generation of scientific insights relating molecular events to patient health are supported by a comprehensive infrastructure: – *The barrier to entry for clinical researchers to do sophisticated molecular medicine is greatly reduced* • Patient privacy and confidentiality is ensured in a consistent and well-vetted manner

caBIG support for basic research lies within the TBPT, ICR, and the recently initiated *In Vivo* Imaging workspaces.

4.1 Tissue Banking and Pathology Tools

Most clinical cancer research is predicated on the availability of a biospecimens of the right type and quantity; this is the starting point for subsequent genomic and proteomic experimentation. In many cancer centers, managing and maintaining the growing number of biospecimens available in frozen and/or paraffin blocked forms presents a problem. Not only are these systems relied on for the logistics of localizing individual specimens within repositories such as large freezer banks, they also are often required to ensure that any specimen use complies with the subject's consent agreement local IRB guidance, and the Health Insurance Portability And Accountability Act (HIPAA) for HIPAA-covered entities.

Systems for managing biospecimens have been developed independently throughout the cancer center community, making interoperability a significant challenge. The TBPT workspace seeks to bind those systems together into a unified resource via a shared informatics infrastructure that enables cross-center interoperability, using the tools shown in the table below.[2] Three software systems are being designed or modified to facilitate integration and access to information from geographically separate areas.

[2] The table below, and those in subsequent sections, identifies the name of the tools for this workspace, its caBIG developer and caBIG principal adopters. It also provides a brief description of the tools.

Table 4-2. TBT Tools

caTISSUE Core	Cancer Text Information Extraction System (caTIES)	caTISSUE Clinical Annotation Engine (CAE)
Provides biospecimen inventory, tracking, and basic annotation capabilities for biospecimen resource facilities. Is developing the foundation software object model, defining core common data elements, and implementing basic functionality of the more expansive caTISSUE system that will be developed in later phases.	A locator to tissue resources via the extraction of coded information from free text surgical pathology reports (SPRs), is using controlled terminologies to populate caBIG-compliant data structures. Provides researchers with the ability to query, browse, and acquire annotated tissue data and physical material across a network.	A Web-based user interface for standards-based manual annotation of biospecimens with clinical information. Supports importing structured data from clinical information systems such as anatomic pathology laboratory systems (APLIS), cancer tumor registries, and clinical pathology laboratory systems. Allows the integration of annotations from multiple sources within the cancer centers, providing a complete picture of a patient's disease.

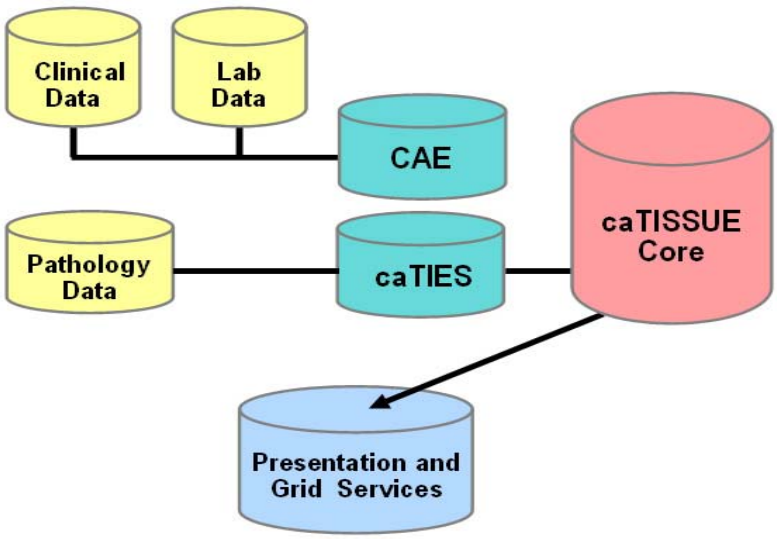

Figure 4-1. TBPT Tools (figure provided by caBIG)

Deployment of these tools will occur in institutions that have varying levels of sophistication in managing biospecimens. Three strategies are envisioned:

1. **Institutions without existing systems:** caTISSUE (the Cancer Tissue Database) Core, including the database backend and Web application, may be deployed and used as a complete application, useful in locations that have been using spreadsheets and similar ad hoc tools to manage their data.
2. **Institutions with existing systems that wish to migrate fully.** An interface that will extract, transform, and load the legacy data into the caTISSUE Core database will be utilized.
3. **Institutions with existing systems who do not wish to migrate.** Adapters will be created that present data to the outside world exactly as caTISSUE Core would, by wrapping interfaces around the existing systems and data elements.

In addition to software development, the TBPT workspace works to develop a coordinated strategy for patient de-identification that is comprehensive enough to cover the major requirements of institutions participating within caBIG. In addition, the workspace is discussing the adoption of currently available standards and mapping out a strategy that ensures that TBPT-specific extensions will be accommodated.

4.1.1 Integrative Cancer Research (ICR)

The ICR workspace is responsible for coordinating the development of bioinformatics tools and standards for the basic and clinical research communities. Although software adopters are contracted to use the tools on specific research projects, many adopters have begun to use the tools more broadly, including for the analysis of non-cancer data. In addition, a number of other groups are using the ICR tools and providing additional feedback. The ICR is organized into five areas, each of which addresses a key technical challenge within bioinformatics.

4.1.1.1 Microarray Repositories

A large obstacle in the maturation of high-throughput DNA microarray technology has been the lack of operating procedures and standards that allow data to be shared beyond the lab of origin. Differences in platform technology, probe selection, and sample preparation all hamper interoperability. Although significant progress has been made with the emergence of the Minimum Information About a Microarray Experiment (MIAME) and MAGE-ML standards for microarray metadata, moving the community to widespread adoption remains a challenge.

The ICR workspace has recognized the need for the caBIG community to be a leader in establishing the use of standardized microarray databases for cancer research. The Microarray Repositories SIG focuses on developing gene expression data storage systems that are MIAME and MAGE-ML compliant.

Table 4-3. Microarray Repositories

caArray	**Developer: NCICB Adopters: Georgetown, Wistar, New York**
A microarray database with open interfaces, strong security, and a user interface that is designed to make standardized annotations (MIAME 1.1) as easy as possible. Allows day-to-day management and analysis of microarray data, facilitates data exchange between research centers and NCICB, and allows data to be easily migrated to the central caArray (Cancer Array Informatics Project) database at the NCI when the data is published. Includes support for MAGE-ML import and export, utilities for the submission and retrieval of Affymetrix and GenePix native file formats, and accessibility through a Microarray and Gene Expression Object Model (MAGE-OM) application programmers interface.	
NCI-60 Data Sharing	**Developer: CCR Adopter: Sloan**
Provides access to genomic and proteomic databases that have assessed potential molecular targets for > 100,000 chemical compounds on the 60 diverse human cancer cell lines used by the NCI Developmental Therapeutics Program.	
Zebrafish microarray	**Developer: Thomas Jefferson Adopter: Sloan**
Provides datasets from a repository of sharable data generated by a custom microarray using the commercial Zebrafish oligo library for use by those working on zebrafish as a model organism.	

4.1.1.2 Genome Annotation

Advances in high-throughput experimental techniques have dramatically increased the ability to specify the order of nucleotides within a fragment of DNA or amino acids within a protein. However, this sequence information often has only marginal scientific value in and of itself. The real scientific value comes when the sequence proper is further enriched via annotation. It is through annotation that the sequence is placed in its biological context and its function in health and disease is understood.

Acknowledging that the value of sequence data is only as good as the quality of the associated annotation, the ICR workspace has dedicated resources to develop both data resources that set the standard for annotation quality and software tools that speed and improve the annotation process. One of the tools, GoMiner, is shown in the figure below.

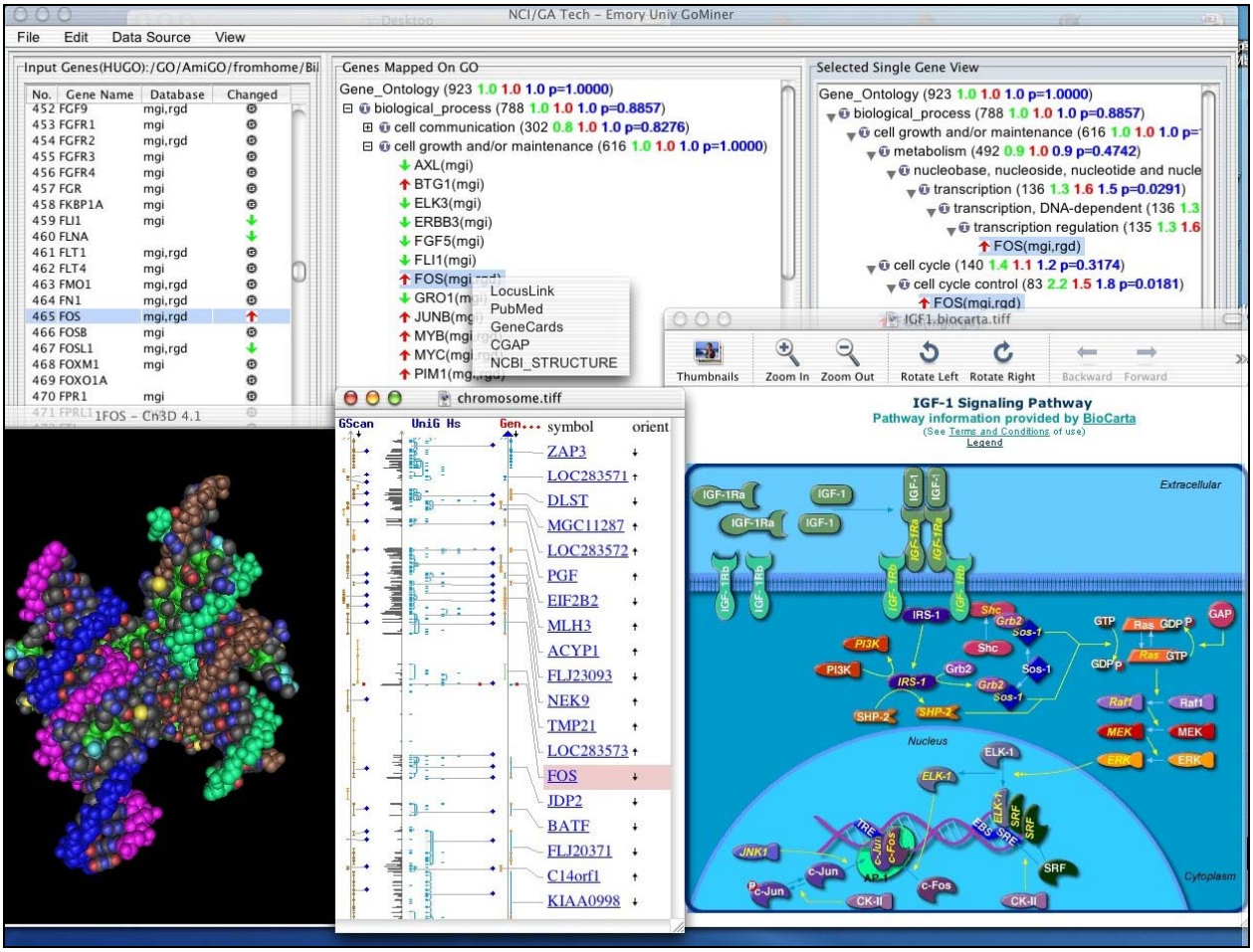

Figure 4-2. GoMiner Screen Shot

Table 4-4. Genome Annotation Tools

Cancer Molecular Pages	**Developer: Burnham**
Database and automated annotation system that combines: 1) automated computer-based annotations; 2) automated data collection from experimental stations; and 3) Web-based visualization tools.	
Function Express	**Developer: Wash U Adopter: Wistar**
Supports the annotation probes on microarrays using publicly available biomedical databases, and provides for the automatic updating of these annotations on a regular basis. Annotation data can be viewed using a Web-based query interface or may be accessed directly by computer programs.	
GOMiner	**Developer: NCI-CCR Adopter: Wistar**
A tool utilizing the Gene Ontology to support the biological interpretation of data from gene expression microarrays and other high-throughput sources. Identifies the biologically coherent functional categories in the frequency of dozens or hundreds of genes that differ in expression between samples. (http://discover.nci.nih.gov/gominer/)	
HapMap	**Developer: Cold Spring Harbor Adopters: Sloan, Wistar**
A database of human single nucleotide polymorphisms, their genotypes, and the linkage disequilibrium relationships among them. (http://www.hapmap.org/)	
Protein Information Resource (PIR)	**Developer: Georgetown Adopter: Upenn**
An annotated protein database that contains more than 283,000 sequences covering the entire taxonomic range. Supports the identification and interpretation of protein sequence information, and assists in the propagation and standardization of protein annotation. (http://pir.georgetown.edu/)	
SEED	**Developer: Holden Adopter: Georgetown**
A framework that supports the peer-to-peer annotation of genomes. A tool for the creation and querying of genomic annotations. Supports the comparative analysis of functional subsystems across organisms; 200 subsystems (primarily prokaryotic, but extending to eukaryotic) have been created to date that cover much of core metabolism.	
Vertebrate PromoterDB (VPD)	**Developer: Cold Spring Harbor Adopters: Sloan, Wistar**
A curated database for vertebrate transcription factor binding sites and their corresponding regulatory regions.	

4.1.1.3 Proteomics

Alternative splicing and post-translational modification processes imply that gene expression analyses, while necessary, are not sufficient for characterizing protein function in cells and tissues of interest. High-throughput proteomic experimental techniques such as mass spectroscopy enable the large-scale study of protein structure and function, but generate extremely large datasets that present significant challenges in data management and analysis. The ICR workspace is developing data management and computational tools for the storage and analysis of both low- and high-throughput proteomic data for the caBIG community.

Table 4-5. Proteomics Tools

Proteomics Laboratory Information Management System (LIMS)	**Developer: Fox Chase Adopter: Moffitt**
A tool for managing proteomic data. Initial efforts focus on tracking the lab processes relevant to 2D gel electrophoresis, but the schema will support the addition of new data types as they emerge.	
Q5	**Developer: Dartmouth Adopter: Oregon Health**
A tool for the probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. Currently implemented in MatLab, Q5 is being ported to a caBIG.	
Proteomics	**Developer: Duke Adopters: Oregon Health, Upenn**
An open source software language for statistical computing and graphics. A tool for post-processing mass spectrometry data; leverages the open source R statistical package.	

4.1.1.4 Pathways

Disease states can be triggered by dysfunction of signaling, metabolic, or transcriptional network behavior. Often there is not a single gene or a single protein cause, but rather a convergence of multiple factors. Understanding how these pathways operate in health and disease involves understanding complex interactions among large numbers of proteins and nucleotide sequences. The highly nonlinear nature of these networks generally makes it extremely problematic to infer the overall network behavior from a static network description. For the caBIG community, ICR is taking the leading role in supporting the understanding of biochemical networks by developing tools and data resources that 1) standardize the representation biochemical networks to promote the sharing of pathway data across researchers; 2) provide high-quality annotation of pathways; and 3) provide computational tools for analyses of networks.

Table 4-6. Pathways Tools

Pathways Tool Development	**Developer: Sloan Adopter: Oregon Health**
A suite of tools for visualizing and interacting with information in the context of biological pathways. These tools include: BioPAX (Biological Pathway Data Exchange Format); a common exchange format for pathways data; cPath, a database focused on protein-protein interactions; and Cytoscape, a bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other statistical data.	
Quantitative Pathway Analysis in Cancer (QPACA)	**Developer: UCSF Adopter: Oregon Health**
A pathway modeling and analysis system that supports the exploration of quantitative biological data in the context of a pathway description. Supports the visualization and computational analysis of pathways.	
Reactome (GKB) Data	**Developers: Cold Spring, Panther Informatics Adopter: Sloan**
A curated database of fundamental biological pathways in humans that uses strict rules of assertion and evidence tracking to ensure a consistent, high-quality product.	

4.1.1.5 Data Analysis and Statistical Tools

High-throughput genomic and proteomic experimental techniques are capable of producing extremely large volumes of data that characterize a large number of distinct genes or proteins. Unfortunately, despite this wealth of data, the number of replicates for a given gene or protein is often quite low. This very high-dimensional data, with a potentially low number of replicates, presents significant statistical and analysis challenges. Using techniques from the statistics, machine learning, and pattern recognition disciplines, the ICR workspace is developing tools to support the discovery of potentially important patterns within datasets of this nature.

Table 4-7. Data Analysis and Statistical Tools

Distance-Weighted Discrimination (DWD)	**Developer: UNC-Lineberger Adopter: Wistar**
Performs statistical corrections to reduce systematic biases caused by differences in microarray platforms, batches of microarrays, or sources of RNA. In MATLAB, to be ported to caBIG.	
GenePattern	**Developer: MIT Broad Adopter: New York**
An analysis platform for genomics research that contains many of the current computational methods used to analyze genomic data; extensible via integration of custom algorithms and visualization tools. Supports the construction of analytical pipelines that use a combination of tools. (http://www.broad.mit.edu/cancer/software/genepattern/)	
Magellan	**Developer: UCSF Adopter: Upenn**
A Web-based system that allows the upload, storage, and analysis of multivariate data and textual or numerical annotations. Is focused on being user-friendly to allow bench biologists to perform complex analyses on heterogeneous data.	
TrAPSS	**Developer: U Iowa Holden Adopter: Wistar**
Tools for searching for the genetic mutation or mutations that cause a defect or disease. Supports: 1) the creation and prioritization of a large candidate gene list, 2) the selection, ordering, and managing of primer pairs, and 3) SSCP assay results.	
Visual and Statistical Data Analyzer (VISDA)	**Developer: Georgetown Adopter: Wistar**
Analysis tool set for discovering patterns in high-dimensional data sets. Supports multivariate cluster modeling and advanced data visualization. Is being ported from MATLAB to open source.	

4.2 *In Vivo* Imaging Tools

The *In Vivo* Imaging workspace was established in October, 2005, to establish standards and interoperable information management systems for imaging datasets, to develop standards and software tools for image annotation, and to develop powerful computational approaches for image processing and analysis. Several projects have been identified:

- The development of an image mark-up standard and associated open source and free annotation creation and display tools, as well as protocols and capabilities to use these tools in a standardized manner on a variety of displays. To date, no accepted standard markup schema have been developed, nor are standard authoring or display tools available.
- The development of a standards-based vocabulary for radiology and allied imaging fields. Early expectations are that a low percentage of the required concepts exist in current caBIG standards.
- The creation of reference datasets for imaging. This entails creation of a uniform set of reference imaging datasets, across modalities, diseases, and organs in order to baseline and help test and develop new imaging systems and technology.
- Development of tools that can de-identify data while still allowing multiple images acquired over time or with different technologies to remain linked, and keeping all of these associated with de-identified imaging reports and other associated metadata for that same individual.
- Development of standards that enable rapid comparison and co-registration of images generated by multiple modalities.
- Development of an extensible imaging platform for development and testing and research and potentially clinical use of a variety of algorithms and applications
- The development of natural language processing tools to extract information from radiology reports.
- Development of imaging standards for small animal studies. Currently few standards apply to the growing field of small animal imaging. As these studies are increasingly used to speed the development of novel pharmaceuticals and to inform early clinical trials, such standards are becoming crucial.

Initial efforts of the Imaging workspace will focus on developing the Reference Research Workbench that supports viewing and annotation of images retrieved from a distributed database.

5. Clinical Trial Management

caBIG is working to harness new developments in molecular discovery by coupling them with clinical practice to expand the pace of biomedical discovery. To do this, caBIG is developing a comprehensive set of modular, interoperable, standards-based software applications to support all aspects of initiating and executing a clinical trial.

For 2006, the Clinical Trials Management Systems Workspace (CTMS WS) is focusing on the Cancer Adverse Event Reporting System (caAERS), a patient study calendar system, a laboratory data hub, all of which will be able to share data with the Cancer Central Clinical Database (C3D). The CTMS WS is also addressing making legacy clinical trials management systems compatible with caBIG. A related project is Firebird, a tool to automate and centralize the 1572 investigator registration for both the FDA and the pharmaceutical industry. Firebird is part of an initiative known as the Clinical Research Information Exchange (CRIX), In addition to facilitating new discoveries in the field of medicine, caBIG will make the process of clinical trials less cumbersome for both researchers and participants. To do this, caBIG is encouraging the rapid adoption of existing quality tools where these can be made compatible.

The table below illustrates a few of the basic improvements to the clinical trials management process that will be enabled as caBIG is fully implemented. The key benefits arise from using standardized data structures and tools, because investigators will spend less time setting up and managing trials. Recognition and management of unexpected trends and adverse events will also become more automatic through standardized reporting and interfaces to laboratory results.

Note that caBIG is not developing all the clinical trials applications de novo; they are encouraging the rapid adoption of tools that already exist and where they can be made compatible. For example, the Cancer Central Clinical Database tool is available for use. Because it is based on an Oracle platform, it is not free, but it does offer a full suite of clinical trials management tools that can be interfaced to caBIG as caBIG evolves.

caBIG is working with the Clinical Data Standards Interface Consortium (CDISC), which is developing standards for electronic acquisition, exchange, submission, and archiving of clinical trials data and metadata for medical and biopharmaceutical product development.[6] By working with CDISC, caBIG is ensuring that its products will be compatible with those being developed for the pharmaceutical industry. This will enhance caBIG's attractiveness for participating institutions that conduct clinical trials for both the NIH and industry. It also will allow the FDA to focus on developing and managing a single data interface to receive electronic trials data.

As part of this effort, caBIG is developing the Biomedical Research Integrated Domain Group Model (BRIDG) model. BRIDG's goal is defining a computable protocol representation that supports the entire life-cycle of clinical trials protocol. Ultimately, tools that use these protocol representations will be developed.[7] BRIDG is developing a formal model of the shared semantics of regulated clinical trials research. This formal model intended to support much more robust interoperability across caBIG-compatible systems.

Consider the representative functional tasks in the table below. They illustrate some of the basic improvements to the clinical trials management process that will be enabled as caBIG is fully implemented. There are other benefits that may be realized; space does not permit including them all. The key benefits arise from having standardized data structures and toolkits because the investigators

will have to spend much less time in setting up and managing the trials administratively. Recognition and management of adverse events will also be greatly enabled by standardized reporting and interfaces to laboratory results. The following scenario illustrates the current approach and how it might be changed under caBIG:

> A principal investigator (PI) is working on a proposal for a new clinical trial. She reviews previous protocols reported in the literature, but finds that many of them are summarized and do not document the entire set of considerations and activities that are required. Reviewing protocols from within the institution, the investigator discovers that they are in many different formats and are difficult to compare. She spends many hours documenting her protocol and then moves on to develop her budget. Developing her budget requires a number of iterations because it is not clear what will be covered by payers and what formats will be required for the various budgeting tools, and she has no useful examples of previous trials for comparison. In particular, it is difficult to predict accrual because the disease burden in her network for the relevant study topic is not well known.

Once caBIG is available, she will be able to use the grid to identify relevant protocols for comparison and analysis. They all will be delivered in a structured form, with associated IRB information, so that they can be compared and modified. The specifications for necessary laboratory data also will come in standardized form, so it will be relatively straightforward to select relevant tests, price them, and determine how they will affect the overall budget. By running a query on participating networks, it will be possible for her to determine the relevant disease burden for the study topic and to take into account the relative success rates for recruiting participants of that type for previous studies. She will be able to create a more accurate budget and submit it for review in a standardized form.

Table 5-1. Clinical Trials Tasks and Benefits

Representative Tasks	Without caBIG	With caBIG	Benefit
1. Formulating a protocol	• Draw upon available protocols at own site or from previous trials • Write protocol in text document or spreadsheet • Distribute changes via email, etc.	• Use caBIG to locate protocols that have already been developed for similar trials • Use caBIG to support complete PI authoring, custom report generation, data mining, results repository, and statistical analysis components	• Integrated support for the development of a clinical trial protocol • Electronic results repository, integrated and standardized data structures making data more widely useful to the cancer community, capabilities for various analyses • Ability to compare results across sites, trials
2. Strategically managing a clinical trial (starting and stopping arms of a study, etc.)	• Analyze results as they become available from paper forms • Identify successful interventions • Identify adverse events • Generate all required adverse drug event (ADE) reports (each in a different format, with different reporting rules, etc.)	• Automated collection of laboratory data in structured format as soon as the laboratory results are completed • Automated, real time data available on adverse events • Automated notification to study leadership as ADEs are detected • Generate all required reports to IRBs, FDA, sponsors, etc., using a standardized, automated interface.. • Analysis capabilities available, can notify the PI of unexpected findings	• Automated receipt of lab data allows more rapid identification of adverse trends, perhaps before the participant develops overt symptoms • Standardized data representations allow for the development of ADE detection algorithms that can be used across multiple sites and trials • Automated ADE report generation to regulatory authorities
3. Manage interactions with payers for clinical trial	• Develop a budget • Develop a study calendar • Review charges to determine what can be billed to payers	• Automated identification of billable events • Automated tracking of payers by participant	• Integrated toolkit for financial management that can be worked out once for all clinical trials with the institution's financial services group. • Ability to build standardized interfaces to billing systems
4. Recruit patients into the clinical trial	• Send notices to participating clinical centers, seeking participants • Receive paperwork for candidate participants • Correspond with investigators concerning eligibility of specific potential participants	• Access online information regarding potential matches across an entire research network electronically, with structured eligibility criteria interfacing electronically with electronic health records • Eligibility criteria validation enabled by workflow tools	• Much more rapid identification and screening of potential participants • Access to a much more diverse participant pool • Access to structured electronic information about potential participants that can be loaded immediately into clinical trial management tools

5.1 Clinical Trials Management Components

Table 5-2. Clinical Trials Management Tools

Clinical Trials Management Tools	
Name	**Description**
Adverse Events Module	The adverse events module will be a comprehensive, highly automated, modular solution for monitoring and managing adverse events that generates reports for submission to internal and external regulatory monitoring organizations.
Laboratory Interface Module	A laboratory interface module will allow the exchange of laboratory data across multiple systems and in multiple formats that have been harmonized into common vocabularies and exchanged via standard message formats.
CDUS/CTMS Reporting Module	A regulatory reporting interface module will be developed to submit data electronically to NCI's Clinical Data Update System (CDUS) and CTMS. This module will capture relevant data from multiple systems and in multiple formats and translate them into the required formats. The process will be automated as much as possible to improve workflow and reduce manual operations. Additionally, this module will allow the retention of data that is lost under current reporting mechanisms in order to facilitate internal analyses of ongoing studies.
Financial/ Billing Module	A number of financial and billing modules are being developed or investigated for development, based on community priorities. These include study calendars, budget development and review, contract review, billing, and study close-out.
Structured Protocol Representation	A method is being developed for documenting protocols so that they may be more readily created, managed, and reported. The method will be compatible with CDISC.
Firebird	An FDA-accepted clinical trials investigator 1572 registry that is also accepted by the pharmaceutical industry is being developed.
caMATCH (pilot)	A patient-centric online clinical trials matching program in a pilot project for breast cancer patients in the San Francisco Bay Area, caMATCH is It allows much better matching of patient clinical conditions to candidate clinical trials, and patients can search for trials themselves using their automated personal health records (PHR). Phase II pilot will create a "middleware" service that can accept patient data from an electronic health record or other digital source and map back to eligibility criteria from protocols.
JANUS	JANUS is a database of key clinical trial information and documents.

6. caBIG Informatics Approach

6.1 Guiding Principles in Infrastructure Development

The caBIG community has embraced four guiding principles to help shape the development of the information infrastructure[8]:

- **Open source software:** Software source-code developed under caBIG funding is freely available. Importantly, the caBIG licensing agreement maintains a pathway to commercialization; interested industry partners can utilize caBIG components in commercial products
- **Open access to data:** In accord with the NIH Data Sharing Policy[9], caBIG emphasizes sharing data for present and future studies in a manner that is compliant with regulatory guidelines and ensures proper attribution for the lab of origin
- **Open development:** Software development occurs in an open and transparent manner, with design choices being vetted in public electronic forums (http://gforge.nci.nih.gov/) that solicit feedback from a broad base of potential developers and end-users
- **Federated systems:** Data and services are geographically distributed across caBIG participants, allowing a level of autonomy, data-ownership, and responsibility to be maintained by the originating organization

A number of operating constraints have emerged, such as:

- The Model Driven Architecture (MDA)[3] approach to software development has been embraced as the recommended development paradigm. Object-oriented representations of data and services are required to achieve full caBIG compatibility.

6.2 Governing the Development Process

Oversight of NCI-funded software development within caBIG strives for a balance between central management and local initiative. The nine workspaces are charged with coordinating software being developed within their areas. Each software component developed in a domain workspace has associated with it a primary developer and a set of one or more formal adopters. The developer-adopter(s) teams are working through a process of specifying user requirements, developing the code, performing initial testing and installation, and gathering end-user feedback. As prescribed by the MDA approach, and required by the caBIG developer contracts, key portions of this process are documented electronically so that they can exchange artifacts (e.g., a document specifying user-requirements, or an image characterizing the implemented data model). A general contractor oversees the operational aspects of the caBIG – including the activities of workspaces and working groups--and is directly accountable to NCICB leadership.

[3] MDA is further discussed in the caCORE section, below. For more detail, see The Object Management Group's Model Driven Architecture Guide at http://www.omg.org/docs/omg/03-06-01.pdf

Working closely with the nine caBIG workspaces, the caBIG oversight board, comprising the leadership of NCICB and the general contractor, provides strategic and programmatic guidance and is the final arbiter of all caBIG pilot activities and decisions.

6.2.1 Vocabulary and Common Data Elements

There are two key technical challenges to developing an information infrastructure that enables broad interoperability of disparate data types. The first challenge involves the data itself. Achieving the statistical power necessary to correlate molecular and clinical variables requires aggregating data that has been gathered over multiple research activities, likely carried out at different cancer centers. Differences in terminology, semantics, and underlying data models need to be resolved before data from different sources can be used effectively. The lack of data interoperability within, and across, basic research and clinical communities and institutions can force researchers to invest substantial effort into harmonizing the data. The Vocabulary and Common Data Elements workspace focuses on developing tools to streamlining this process and promote interoperability.

6.2.2 caCORE (the Cancer Common Ontologic Representation Environment)

The Cancer Common Ontologic Representation Environment (caCORE) provides a set of tools for developing data resources that enable their contents to be more easily shared within an institution and across institutional boundaries. Adopters of the caCORE framework annotate their data using a controlled vocabulary (the NCI Thesaurus) so that humans can more quickly understand how to interpret the data and software can easily be developed to allow interoperability between systems.

Historically, assembling multiple data resources into a single, *integrated* resource has been a difficult problem because of several types of heterogeneity:

1. **Systemic heterogeneity:** The underlying database systems used to store the data may not be compatible.
2. **Syntactic heterogeneity:** Database developers often choose different representations for the same data element. For example, sequence data can be stored in multiple formats (such as FASTA, ALN, or Clustal).
3. **Structural heterogeneity:** The organization of data elements may differ across the underlying databases. As a simple example, should author records reference books or book records reference authors (or both)?
4. **Semantic heterogeneity:** Different database developers may use the same term to refer to different concepts (e.g., is a nail a body part or a surgical supply?) or use different terms to refer to the same concept (e.g., myocardial infarction vs. heart attack).

To integrate multiple data resources, an integration engineer first determines which data resources to integrate, based on prior knowledge or social networking. He must then figure out how to pull data out of the systems responsible for managing the data (i.e., he solves the first problem). He then identifies semantic overlap across the systems (solving the fourth problem). Finally, he writes code to perform structural and syntactic transformations.

The caCORE toolkit reduces the barriers to integrating data resources. The first tool provides a framework for modeling data resources that enables the structure of the data (its *model*) to be exported in a common format that is easily understood by the remaining caCORE tools.

The second tool helps the developers annotate the model with terms drawn from a common vocabulary, the NCI Thesaurus. This annotation minimizes semantic heterogeneity by associating with each data element a unique concept identifier that describes the element. The third caCORE tool uploads the model into a metadata registry so that other caBIG participants can easily locate data resources relevant to their research.

The last tool generates a common application programming interface (API) to the data resource. When developing a new data resource, this tool also can automatically generate a database system and create the software for extracting data from the database according to the API. When publishing an existing data resource, a software developer must determine how to connect the API to the database.

Using these tools, more scientists will be able to augment their own data with data resources published to the grid. The steps required for this integration are reduced to the following:

1. Search for relevant data resources using the metadata registry.
2. Determine which data elements are relevant based on the semantic annotations.
3. Retrieve these elements using the published API.
4. Perform any necessary structural or syntactic transformations.

The caCORE toolkit fulfills its goals of reducing the cognitive effort needed to interpret data and lowering the barriers to developing software that pulls data from multiple data resources. This is not to say that the toolkit is a panacea for data integration—for example, there is an inherent trade-off between the ease with which a model can be annotated and the extent to which subtle variations in meaning can be communicated. Regardless, caCORE successfully reduces the amount of programming expected of informaticists supporting clinical and biologic research.

6.3 Architecture

caBIG utilizes a service-oriented architectural approach and defines the set of core capabilities needed to support the sharing of informational and computational resources among researchers, scientists, and healthcare professionals. The core capabilities can be divided into those concerning service identification and invocation, those concerning information semantics, and those concerning security. In addition to the core capabilities, caBIG currently has six nodes that supply a variety of cancer-related capabilities (e.g., Microarray data services).

The caBIG architecture is based on the Globus toolkit, an open source technology that allows users to share computing power, information, and other tools securely across geographic areas without sacrificing local autonomy. This means that while the caBIG information and services are available for other to utilize, the local authority (e.g., a Cancer Center) maintains governance over the information and services. This federated approach allows the caBIG community to share information without disrupting the local operations of the participants.

6.4 caBIG Compatibility

The caBIG compatibility guidelines[10] define the processes that must be followed and artifacts that must be produced for a node to be considered caBIG compatible. Bronze, silver, and gold levels of compatibility are envisioned, with higher levels having improved interoperability characteristics.

Currently caBIG has defined the processes, procedures, and artifacts needed for a node to be certified as compatible to the bronze level. Four elements are needed:

1. **Interface Integration:** Programmatic access to data from external resources must be available
2. **Vocabulary and Terminology:** Publicly accessible, controlled vocabularies must be used; all terms must meet the VCDE workspace guidelines.
3. **Data Elements:** Detailed data element definitions that are built using controlled terminologies must be used, with that metadata stored electronically and available separately.
4. **Information Models:** A diagrammatic representation of the information model must be available electronically.

Currently, tools developed within caBIG™ are internally reviewed and approved for silver-level compliance. NCICB is exploring the creation of a formal verification process for certifying the compatibility of applications developed outside the management and oversight of the caBIG™ program. A preliminary process, for Bronze compatibility, is outlined on the caBIG™ Web site at https://caBIG.nci.nih.gov/guidelines_documentation. The caBIG licensing agreement has been specifically designed to ensure that industry partners can actively participate within the community. The Bronze Compatibility Certification was set up to authenticate requirements from projects within the workspace in order for those projects to earn the license to use the caBIG trademark as a selling point.

6.5 caBIG: Security and Privacy

Translational research spans clinical and basic-science communities. Human research volunteers participate across this spectrum, assuming roles such as those of : patients in clinical care, patients enrolled in clinical trials, and volunteer for human subject research. Importantly, different regulatory guidelines are operative for research volunteers in these different roles. For caBIG, the three key regulatory concerns are:

- The Health Information Portability and Accountability Act (HIPAA) and associated state laws
- The Common Rule for Human Subjects Research
- The FDA Rule for Human Subjects Research
- The FDA's 21 CFR Part 11
- Local guidelines and interpretations of the IRB

A recent report from the American Hospital Association asserted that state laws are a much greater barrier to clinical information sharing than HIPAA itself. Therefore, the need for configurable security models is all the more apparent.[11]

In addition to the regulatory issues, wide-spread adoption of team science is highly contingent on ensuring that the lab of origin is confident that its hard-won experimental data will not be used without proper scientific attribution. Many data generators have strong concerns over how attribution will occur. The current reward structure within academic settings, in particular, does not provide powerful incentives for sharing experimental data. If there is any chance whatsoever that further analysis or processing of the data might lead to a publication, investigators tend to be reluctant to release it broadly.

What does resonate with researchers is the ability for them to control access and establish varying degrees of data visibility, such as sharing fully with co-investigators, sharing some pre-publication data with trusted collaborators, and sharing published data freely. The capability to establish such nested circles of trust is a key issue for many potential adopters. The caBIG federated architecture supports this, for example, by not housing data in a centralized repository. Each institution controls the data that it will expose to the grid and to whom the data will be available.

The Architecture workspace, in conjunction with experts on the regulatory and proprietary complexities of data sharing in the Data Sharing and Intellectual Capital workspace, has recently begun the task of developing a strategy for grid security and privacy. The focus to date has been on surveying available options for two key functions:

- **Authentication:** Ensuring that an individual seeking access to grid resources is who they purport to be
- **Authorization:** Determining which resources each individual can has access

Different technical solutions to these key functions exist. Having surveyed these solutions, the caBIG community is in the process of discussing the merits of alternative approaches.

7. Related Efforts

7.1 BIRN (Biomedical Informatics Research Network)

Biomedical Informatics Research Network (BIRN), launched by NIH in 2001, is building an infrastructure of networked, high-performance computers; data integration standards; and other emerging technologies. The BIRN currently consists of three "test bed" projects that are conducting structural and functional studies of neurological disease:

- Function BIRN — studying regional brain dysfunctions related to the progression and treatment of schizophrenia.
- Morphometry BIRN — examining unipolar depression, mild Alzheimer's disease, and mild cognitive impairment.
- Mouse BIRN — studying animal models of multiple sclerosis, schizophrenia, Parkinson's disease, ADHD, Tourette's disorder, and brain cancer.[12]

BIRN is using many of the same technological approaches, such as ontologies and federated data structures distributed across multiple sites, as caBIG.

7.2 National Center for Biological Ontologies

The National Center for Biological Ontologies is an NIH-funded program designed to advance the state of ontology development and implementation in biomedical research. As such, the center will provide useful input for caBIG. There are three projects currently underway that all share the common activities of accessing, editing, and using ontologies to describe biomedical knowledge in their particular domains and of using that ontological knowledge to annotate and analyze biomedical data[13].

7.3 Industry Involvement

NCI welcomes the involvement of commercial and other groups in caBIG, whether they are information technology companies, software or device vendors, pharmaceutical companies, biotechnology companies or non-profit organizations. The level of participation varies; some actually develop caBIG-compatible tools, some participate in work spaces and some are modifying their existing products for caBIG compatibility. Others compete for caBIG contracts. Key standards development organizations such as Health Level 7 (HL7) and Clinical Data Interchange Standards Consortium (CDISC) are also engaged.

Large corporate vendors have already begun to engage in caBIG activities. For example, it was reported at a recent meeting that Cerner Corporation, in support of the Cancer Text Information Extraction System (caTIES) application, has constructed an adaptor that allows data from its co-Path product to be extracted and transformed into a caBIG-compatible format.

7.4 Summary

Acknowledgement of the promise held forth by both translational research and large-scale team science is widespread. But truly realizing this promise involves a sea change in the mindset

of clinicians, researchers, and funding entities working in the life sciences. A number of standard roadblocks are often raised; within the cancer community caBIG is striving to remove them. The hypothetical discussion below illustrates key problems often raised in the research community when data sharing initiatives are discussed and shows how caBIG addresses them to improve the ability of translational science to meet Dr. Zerhouni's requirement for a research framework:

- "IRB, subject consent, and regulatory issues preclude sharing my data. Even with these addressed, there is always one more paper I plan to wring out of that data; I just need the time to get back to it. Attribution issues won't be worked out sufficiently for me to feel comfortable just giving away my hard-won data."
  - caBIG: Getting de-identification applications constructed, vetted, and in use, with very well-tested capabilities, can aid in removing many regulatory concerns. NCI and NCICB have already begun to brief cancer center IRBs about how the grid infrastructure intersects with regulatory issues.
  - Interoperability is enabling a new way of conducting science. Concomitant changes in the reward structure will occur; sharing of data will soon become as career enhancing as wringing out that journal article.
- "Sharing basic science data is not practicable, as no other lab is capable of interpreting and understanding my data to the extent necessary to do good science; I would be doing a disservice to the field to release it broadly."
  - caBIG: Getting the metadata standards surrounding the data right will enable sharing of this data in a manner that affords aggregation of datasets in a well-founded and scientifically meaningful manner.
  - Powerful applications that utilize these standards can be the vehicle for gaining widespread adoption. The use of the standards-based tools that achieve current research and process goals can be a vehicle to drive interoperability.
- "Clinical research is about the patient and biology, all this information technology and computational sciences is confusing and distracting for the clinician and life science researcher, consuming time and resources."
  - caBIG: Life sciences has become increasingly multi-disciplinary and will continue to do so to the benefit of the field. Future life science research is going to be predicated on and tightly tied to the information sciences.
- "The capability to develop information technology has always been available via individual investigator grants; they just need to be embedded in an R01 proposal with a set of specific aims that focus on real science questions."
  - caBIG: Large-scale informatics infrastructure in cancer research is important. It merits a level of commitment that heretofore hasn't existed and requires a different approach to the funding mechanism, switching from grants to contracts.

caBIG is building a cohesive community among the clinical cancer research in which this sea change is occurring. It is as much as about bringing people together to embrace a fundamental change in how science is conducted as it is about developing the enabling technology.

Importantly, the investment made within the cancer community has the potential to be applicable to other domains, and caBIG has already begun to attract widespread interest. The availability

of interoperable clinical, genomic, proteomic, and imaging data could impact virtually every ongoing scientific endeavor within the life sciences. However, key technical challenges and design decisions would arise in attempting to adapt caBIG infrastructure to activities beyond cancer, including:

- How difficult is it to extend the underlying vocabularies and terminologies to new domains?
- Should caBIG be scaled up to cover all of translational science, or should components be replicated and tailored for specific domains?
- Will network performance scale when researchers outside the cancer community begin to push large experimental datasets around the nation?

These and a number of other similar issues merit serious consideration and study as caBIG becomes poised to form the basis for an even grander vision.

Acronyms

APLIS	Anatomic Pathology Laboratory Systems
BioPAX	Biological Pathway Data Exchange Format
BIRN	Biomedical Informatics Research Network
BRIDG	Biomedical Research Integrated Domain Group Model
caAERS	Cancer Adverse Event Reporting System
caARRAY	Cancer Array Informatics Project
CAE	Clinical Annotation Engine
caTIES	Cancer Text Information Extraction System
caTISSUE	Cancer Tissue Database
CDE	Common Data Elements
CDISC	Clinical Data Standards Interface Consortium
CDUS	Clinical Data Update System
FDA	Food and Drug Administration
ICR	Integrative Cancer Research
MAGE-ML	Microarray and Gene Expression Markup Language
MAGE-OM	Microarray and Gene Expression Object Model
MDA	Model Driven Architecture
MIAME	Minimum Information About a Microarray Experiment
NCI	National Cancer Institute
NCICB	National Center for Bioinformatics
NIH	National Institutes of Health
PI	Principal Investigator
PIR	Protein Information Resource
Q5	Name of application in caBIG
VCDE	Vocabularies and Common Data Elements
VISDA	Visual and Statistical Data Analyzer
VPD	Vertebrate PromoterDB

List of References

1. Hörig, H., Marincola, E., and Marincola, FM. Obstacles and opportunities in translational research. *Nature Medicine* 11, 705 - 708 (2005)
2. Cancer Biomedical Informatics Grid (caBIG™) 2006 Strategic Plan (Draft), caBIG - Documents
3. NIH Data Sharing Policy http://grants.nih.gov/grants/policy/data_sharing/
4. The Object Management Group's Model Driven Architecture Guide http://www.omg.org/docs/omg/ (document identification number is: 03-06-01.pdf)
5. NCI-Designated Cancer Centers http://www3.cancer.gov/cancercenters/centerslist.html
6. Buetow KH. Cyberinfrastructure: empowering a "third way" in biomedical research. *Science*. 2005 May 6;308(5723):821-4.
7. Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, Gustafson S, Buetow KH. caCORE: a common infrastructure for cancer informatics. *Bioinformatics*. 2003 Dec 12;19(18):2404-12.
8. Buetow KH. The NCI Center for Bioinformatics (NCICB): building a foundation for in silico biomedical research. *Cancer Invest*. 2004;22(1):117-22.
9. Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, Gustafson S, Buetow KH. caCORE: a common infrastructure for cancer informatics. *Bioinformatics*. 2003 Dec 12;19(18):2404-12.
10. caBIG™ Compatibility Guidelines https://cabig.nci.nih.gov/guidelines_documentation
11. caBIG™ Participants https://cabig.nci.nih.gov/participants/

Contributors

This report was prepared by a team of MITRE staff. Participants included:

- Steven Decker, M.S.
- Jordan Feidler, M.S., Project Lead and lead author
- Brandon Higgs, PhD
- Joseph Mitola, PhD
- Olivia Peters, M.S
- Robert Mikula, M.S.
- Peter Mork, PhD
- Matthew Seguin, M.S.
- Sandra Sinay, R.N., J.D.
- Jean Stanford, Project Manager
- Gary Vecellio, M.S.
- Marion Warwick, M.D.

End Notes

---

[1] Barbara J. Culliton, "Extracting Knowledge From Science: A Conversation With Elias Zerhouni," *Health Affairs* 25 (2006): w94-w103 (published online 9 March 2006; 10.1377/hlthaff.25.w94)

[2] Cancer Biomedical Informatics Grid (caBIG) 2006 Strategic Plan (Draft) https://cabig.nci.nih.gov/working_groups/SP_SLWG/Documents/

[3] Kenneth H. Buetow, Ph.D., "The NCI Center for Bioinformatics (NCICB): Building a Foundation for In Silico Biomedical Research", Cancer Investigation, Vol. 22, No. 1, pp. 117, 2004

[4] NCI-Designated Cancer Centers http://www3.cancer.gov/cancercenters/centerslist.html

[5] Cancer Biomedical Informatics Grid (caBIG) 2006 Strategic Plan (Draft) https://cabig.nci.nih.gov/working_groups/SP_SLWG/Documents/

[6] http://www.cdisc.org/index.html

[7] Douglas Fridsma, M.D., PhD, "The BRIDG Project: Creating A Model Of The Semantics Of Clinical Trials Research", Presented At The caBIG Annual Conference, April 11, 2006.

[8] Cancer Biomedical Informatics Grid (caBIG) 2006 Strategic Plan (Draft) https://cabig.nci.nih.gov/working_groups/SP_SLWG/Documents/

[9] NIH Data Sharing Policy http://grants.nih.gov/grants/policy/data_sharing/

[10] caBIG Compatibility Guidelines https://cabig.nci.nih.gov/guidelines_documentation

[11] *Health Information Exchange Projects What Hospitals and Health Systems Need to Know*, American Hospital Association, May, 2006, p. 14.

[12] http://www.nbirn.net/

[13] http://bioontology.org/index.html