Prepared for:

National Institutes of Health
National Center for Research Resources

caBIG-Plus™ Conceptual View: Beyond Cancer

July 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.

Approved for Public Release; Distribution Unlimited. Document Number 06-1519

MITRE
Center for Enterprise Modernization
McLean, Virginia

National Institutes of Health National Center for Research Resources

The National Institutes of Health National Center for Research Resources (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 http://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 (FFRDC) 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

> "It is the responsibility of those of us involved in today's biomedical research enterprise to translate the remarkable scientific innovations we are witnessing into health gains for the nation."[1] *Elias Zerhouni, M.D., Director, National Institutes of Health*
>
> "We conclude that the clinical research effort in the United States must be seen for what it is—a fragmented cottage industry constituted of multiple stakeholders … with no overarching vision, no cohesive organizational framework, and at times not even a common forum for dialogue or active collaboration. The current poorly articulated and highly compartmentalized components of the existing non-system are inefficient and often redundant. Hence, they diminish effectiveness and increase costs of translating basic research to patient care while often not contributing materially to its safety or efficiency. … Most importantly, this existing infrastructure … is currently functioning on overload. The U.S. capacity to translate basic science into improved health for its population is rapidly being exceeded by the burgeoning scientific opportunities at hand. ***Basically, the 'clinical research grid' is failing***."[2] [Emphasis added.]

Throughout the academic medical community and the healthcare industry, there is a clear imperative to improve the connection between basic research and patient care.[3] To accomplish this objective, the National Institutes of Health (NIH) developed the NIH Roadmap for Medical Research[4] and, in 2004, the National Cancer Institute (NCI) Center for Bioinformatics initiated a pilot project to develop an information infrastructure that enables this connection. The Cancer Biomedical Informatics Grid™ (caBIG™) is an open-source, open-access information network enabling cancer researchers to share tools, data, applications, and technologies according to agreed-on standards and identified needs. caBIG seeks to connect the entire cancer community, from bench scientists to cancer clinicians to reviewers at the Food and Drug Administration.[5]

This is the third of three reports commissioned by the NIH National Center for Research Resources (NCRR) on caBIG and its applications to research. The first report, *caBIG™ Overview*, provides an overview of the current functions and capabilities of caBIG.[6] The second report, *caBIG™: Opportunities and Challenges for Use Beyond Cancer*, describes the technology capabilities and infrastructure required to expand caBIG beyond the cancer community to the entire biomedical community.

The purpose of this third report is to provide a conceptual view of the ways in which caBIG can be expanded to the entire biomedical community.

Some of the emerging technologies for expanding caBIG, such as open source, Service Oriented Architecture (SOA), grid, and Extensible Markup Language (XML), are summarized in Appendix C.

1.1 Statement of the Problem

The biomedical community is facing numerous challenges. We identify some key issues here to frame the discussion of the potential uses of an expanded version of caBIG (also known as caBIG-Plus).

> *Academic Health Centers: Leading Change in the 21st Century* addresses the perception that researchers have to work in multidisciplinary teams to be effective. The book states that "increased coordination and collaboration will be required to meet growing

> demands for rapid improvements in health care and for a greater focus on the types of research that answer questions about what does and does not work. … Basic biomedical research is typically carried out by an individual researcher or team of researchers from the same field [but] some believe that the individual researcher who tries to do it all will flounder, given that the necessary expertise will reside in a team of researchers rather than an individual." It also points out that researchers are going to require a great deal more informatics support, because "research in biomedical fields such as genomics generates immense amounts of data to be analyzed. Correlation of genotypes with phenotypes will require access to longitudinal clinical information and large numbers of patients."[7]

Dr. Zerhouni confirms this point:

> "Future progress in medicine will require quantitative knowledge about the many interconnected networks of molecules that comprise cells and tissues, along with improved insights into how these networks are regulated and interact with each other. … To fully capitalize on the recent sequencing of the human genome and many new discoveries in molecular and cell biology, the research community needs wide access to technologies, data bases and other scientific resources that are more sensitive, more robust and more easily adaptable to researchers' individual needs."[8]

Buneman et al. also confirm this point by stating "Databases are an essential part of the infrastructure of science."[9] The authors add that data must be accumulated in a standardized form, distributed to collaborators throughout a secure environment, integrated with other data sets, analyzed, and curated over many years.

Performing research in teams has become necessary to defray the costs of data collection. By sharing analytical tasks, researchers are able to avoid unnecessary work—not only computational work, but also database setup and integration.

Major challenges exist in integrating the vast quantities of data required for advances in basic science; additional challenges exist in integrating basic science with clinical research. Campbell et al. define clinical research as "research that involves living humans as subjects, is composed of a wide spectrum of research types such as clinical trials, translational research, epidemiological research, health services research and outcomes research."[10] The authors attribute challenges in the clinical research environment to the following:

- Lack of support for clinical research institutions
- Lack of time for young faculty to participate in clinical research, resulting in fewer trained researchers
- Competition from contract research organizations for resources, programs, and staff

Following are the types of strategic adaptations that academic medical centers are pursuing to address these challenges:

- Helping faculty identify funding sources
- Helping researchers identify potential collaborators
- Helping researchers write and review grant applications
- Recruiting research participants

caBIG-Plus capabilities address many of these challenges (e.g., by providing sophisticated collaboration tools and enhanced workflow environments).

1.2 Applicability of caBIG Beyond Cancer

As discussed in the second report, *caBIG™: Opportunities and Challenges for Use Beyond Cancer*,[11] many research requirements are similar across domains, enabling the expanded use of caBIG. The applicability of caBIG to non-cancer activities has been demonstrated. Several institutions are already using caBIG for non-cancer applications or have been involved in the development of caBIG to make its tools appropriate for their non-cancer work. Some tools, including the Cancer Tissue Database (caTISSUE),[12] are already generic enough to be used by non-cancer researchers.

caBIG was designed so that it can be expanded to facilitate biomedical research, aiding researchers and their support staff by streamlining several necessary processes, thereby saving time and money and supporting sound scientific research. By providing an integrated tool set that supports researchers' entire workflow (including management of contracts for supplies and reagents, time accounting, and so on), caBIG could enable researchers to spend much more time conducting their research and much less time on the administrative and data acquisition aspects of their work.

The following capabilities, via a combination of caBIG technology and changes to workflow, could streamline processes that absorb a great deal of researchers' time and resources:

- Integrating the workflow of scientific teams as they move from basic science to identification of intervention targets, to testing of interventions, to trials and studies, to post-market surveillance
- Providing an integrated infrastructure that enables researchers to download data in a standardized format to avoid wasting time writing filters, translating vocabularies, and performing data cleanup work not related to their research
- Enabling the sharing of tools, methods, and technologies so that work can be leveraged across the research community more effectively
- Communicating results rapidly, improving collaboration, and providing better access to reference data sets
- Processing requests for materials (e.g., samples, tissues) across the community more efficiently, thus permitting more convenient access to materials and enabling research not currently plausible because of the expense of collecting many types of data
- Enabling the sharing of protocols, study designs, and other materials so that researchers do not have to "reinvent the wheel" and so that institutional review boards (IRB) can leverage standardized work products and workflows

caBIG currently provides some tools that support these capabilities in the cancer domain, and some of these tools are being used in other research domains. By evaluating the entire research life cycle and applying tools to support researchers in a holistic way, caBIG could evolve into an extremely robust researcher support environment.

1.3 caBIG-Plus Conceptual View

Figure 1-1 presents the caBIG-Plus conceptual view, showing how selected activities could be facilitated in the expanded caBIG environment. These concepts, although not all-inclusive, are intended to help the reader visualize the potential benefits of expanding caBIG.

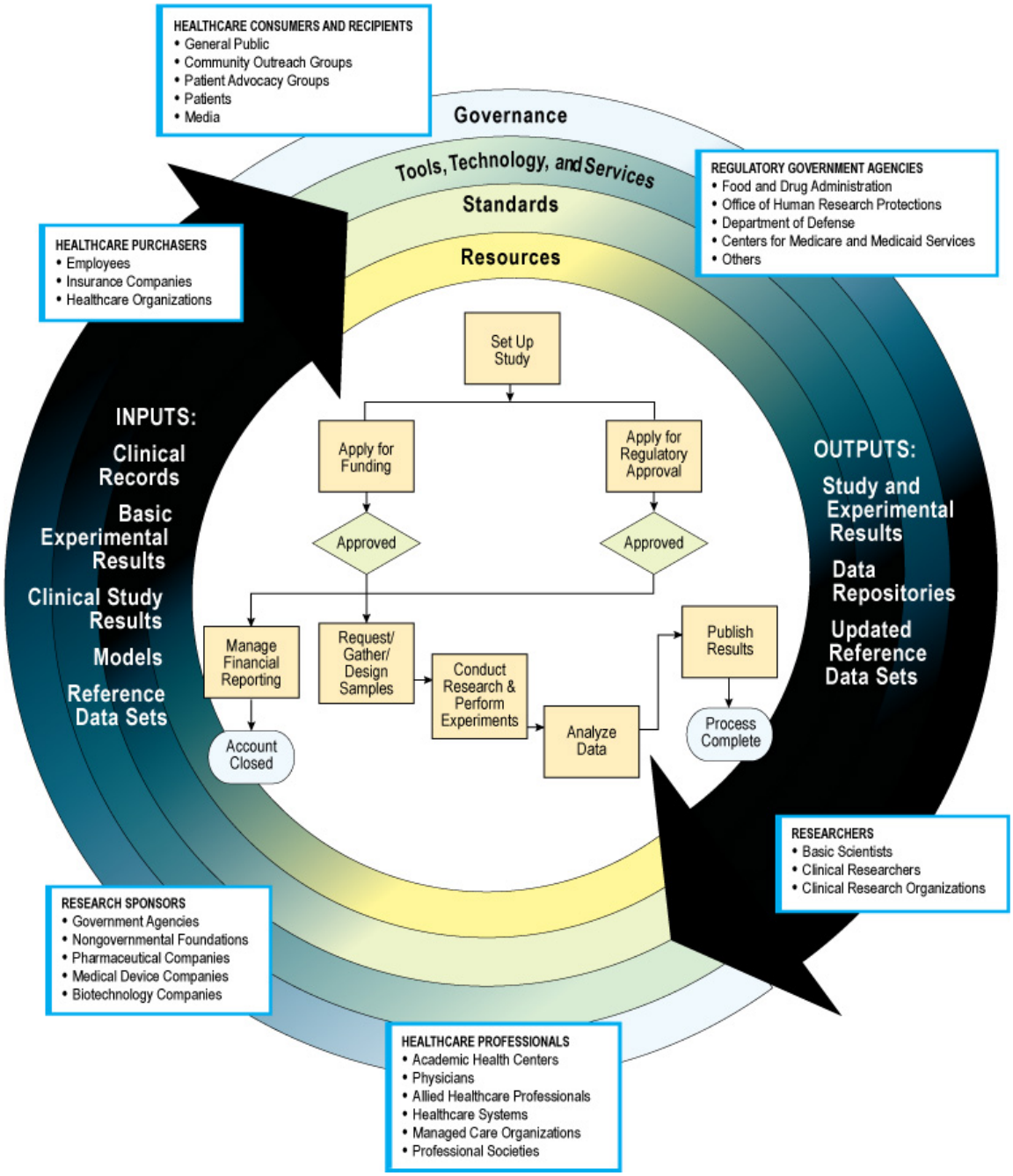

Figure 1-1. High-Level Conceptual View of caBIG-Plus

Table 1-1 explains each component of this conceptual view.

Table 1-1. caBIG-Plus Conceptual View Components

Component	Description
Inputs	• **Clinical Records.** Patient care records, which can be received in an identified, de-identified, or anonymized manner, depending on their intended use. Preferably, these records will be delivered in interoperable electronic form, using standardized vocabularies. • **Basic Experimental Results.** Data sets from researchers within a group or from other groups. Preferably, these results will be brought into the caBIG-Plus environment in standardized form for rapid integration into studies. • **Clinical Study Results.** Data sets that are derived from clinical studies within a group or from other groups. • **Models.** Algorithms and data sets that are used for simulations and other research purposes. • **Reference Data Sets.** Structured collections of data that are used for comparison, calibration, and other purposes during studies.
Stakeholders	• **Healthcare Purchasers.** Will use clinical study results for outcomes research, clinical studies, and other purposes. • **Healthcare Consumers and Recipients.** Will use the caBIG-Plus environment to exchange information on clinical trials and locate information on clinical studies. • **Regulatory Government Agencies.** Will use the caBIG-Plus environment to exchange messages and collaborate with researchers. The agencies will review and approve applications (e.g., for use of experimental drugs) and use specific regulatory information exchange components as research proceeds from approval of an intervention through the post-market surveillance process. • **Researchers.** The primary users of the caBIG-Plus environment, who will exchange data, seek tissues and other materials for experiments, collaborate, obtain access to remote instruments, and so on. • **Healthcare Professionals.** Will use the caBIG-Plus environment to participate in clinical studies and review research results. Healthcare professionals could begin to exchange phenotype and genomic data with basic science researchers through approved data-sharing agreements. • **Research Sponsors.** Will use the caBIG-Plus environment for a wide variety of purposes, including exchanging study data, communicating with regulatory agencies, collaborating with clinicians, and providing new tools and technologies to the industry.
Basic Infrastructure	• **Governance.** Needed to lay the ground rules to create and enforce policies on data sharing, security, interoperability, development processes, and so on. • **Tools, Technology, and Services.** The basic building blocks that enable caBIG-Plus to operate. • **Standards.** The agreed-on means by which technology providers structure their devices, messages, and/or services so that caBIG-Plus participants can use them in an interoperable, predictable manner. • **Resources.** The funding, staff, data, and other components provided to make the caBIG-Plus environment work.

caBIG-Plus could be configured to support a wide range of users and could provide valuable workflow tools to support collaboration needs. However, process capabilities need to be developed to achieve this vision.

2. caBIG Beyond Cancer: Conceptual View of Capabilities

To meet the needs of the entire biomedical research community, additional capabilities and tools need to be included in caBIG. Some of these capabilities and tools would be useful even if caBIG is not expanded for use beyond the cancer research community. Figure 2-1 provides a high-level view of the processes that would be enabled in the caBIG-Plus environment. The purpose of this process flow diagram is to provide a conceptual vision of an enhanced research environment. Bullets that appear in green text indicate current caBIG functionality (though not all tools are in widespread use), and bullets that appear in blue italicized text highlight potential areas of support by caBIG-Plus.

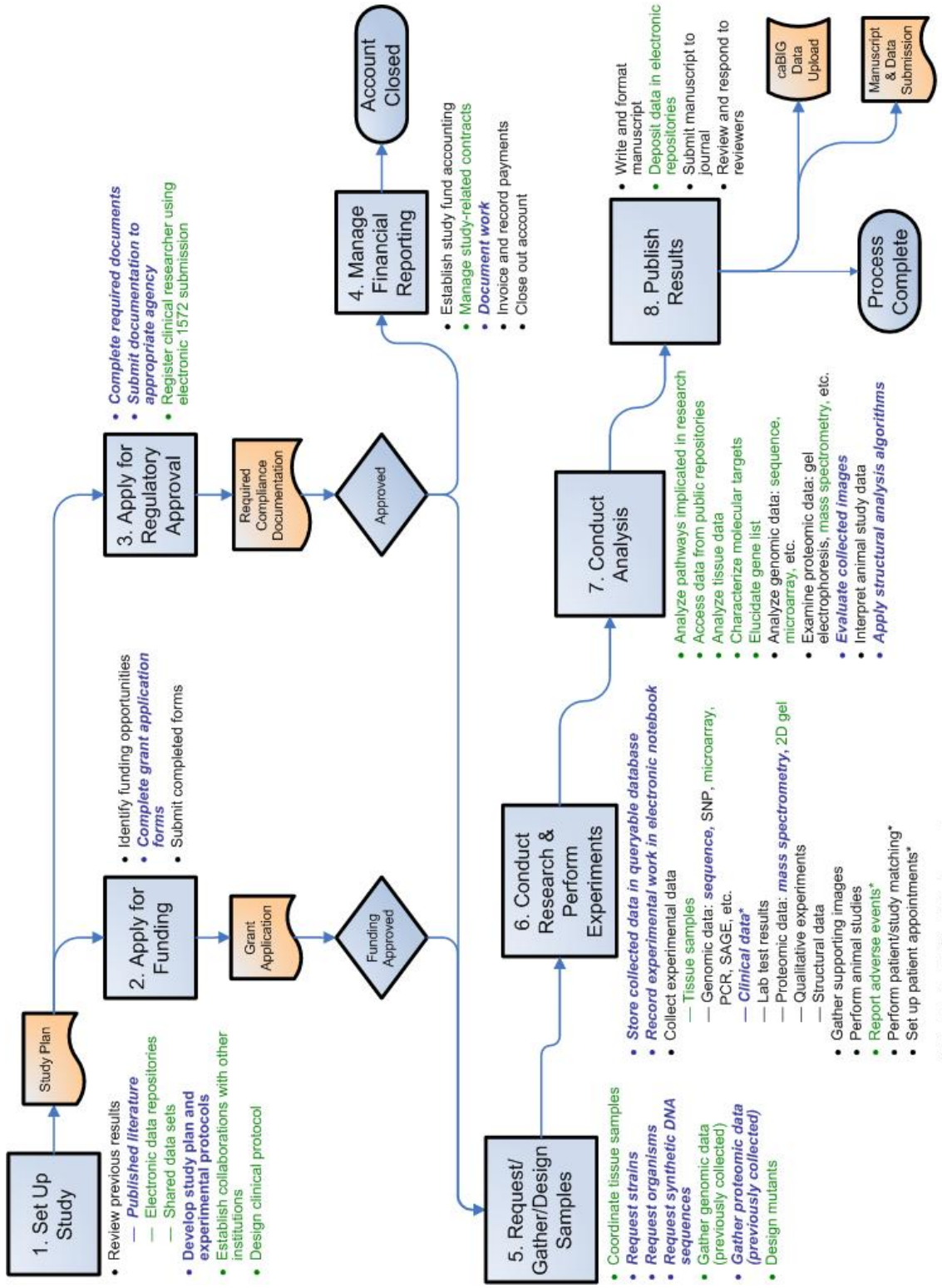

Figure 2-1. caBIG and caBIG-Plus Support for Clinical Research

Table 2-1 highlights the capabilities required in the caBIG-Plus environment to support clinical research.

Table 2-1. caBIG-Plus Capabilities Supporting Clinical Research

1. Set Up Study
Provide an environment for rapid development of a study plan (including protocol, if necessary). Support the researcher in seeking initial data sets and identifying collaborators.
• Provide organized links to the literature and data repositories. These repositories could be searchable and indexed to enable the researcher to rapidly identify appropriate data. Once the appropriate literature and data are identified, the researcher could use a directory to quickly establish a collaborative dialog with the study authors and could subscribe to the literature or data repositories so that any changes to the original data or publications referring to it can be quickly assimilated into future studies. • Provide a link to phenotype and clinical data repositories, as appropriate and approved for access, to enable the researcher to investigate hypotheses, determine the availability of suitable samples or patient cohorts, and so on. In addition, link clinical data to other data collected (e.g., genomic or proteomic information). • Provide a link to executable protocols already in caBIG-Plus–compatible formats. Provide libraries of study plans in a standardized format so that they can be used as templates for new studies. • Provide a protocol workflow that automatically captures study details in executable form and route the study for collaborative review, regulatory review (internal and external), and regulatory approval. • Use a standardized template library, and design a patient consent form (if necessary). Use workflow tools to route the form for review. • Make it easier to identify collaborators, perhaps through the development of a data registry service that enables researchers to find and contact others who are involved in similar research. Ideally, a matching mechanism will be included in caBIG-Plus, enabling users to identify their own interests and peruse the interests of others.
2. Apply for Funding
Support researchers as they identify funding sources and provide a workflow for managing the application, regulatory review, and regulatory approval processes. The workflow would support the overall financial management workflow for researchers whose own institutions do not provide these tools.
• Manage a searchable registry of funding opportunities from all sources. • To streamline the process of applying for funding, provide a standardized application template with electronic form-completion capabilities. Alternatively, maintain a repository of electronic applications for various funding sources to provide quick and easy access for researchers. • Provide workflow tools to manage the flow of the funding application within the institution and to funding agencies or companies. These tools could include a Web service listing all currently available funding opportunities, with links to required application forms. (The service would require an agent that automatically queries many funding sources.) Provide the capability for collaborative applications so that multiple researchers can conveniently submit a single funding application. Develop a way to pass funded applications directly to internal management. • Manage a secure repository where supporting documentation can be maintained for use by funding applicants, sponsors, and regulatory bodies.
3. Apply for Regulatory Approval
Support teams as they go through the internal and external regulatory approval process, providing a one-stop shopping portal to enable collaboration on the workflows necessary to get studies approved.
• Use standardized registries to enable researchers to establish and submit their credentials for conducting the study. • Automate the regulatory approval process. Ideally, a Web service would be implemented to enable the researcher to select a number of options (e.g., human experimentation, work with selected agents) and receive the required regulatory forms and the names of the organizations to which these forms should be submitted. These submissions would preferably be electronic. A further level of improvement could be made if the forms required for internal approval were standardized across institutions to enable their automation as well. Performing researcher registration electronically would be another enhancement (e.g., for clinical researchers using Form 1572, which caBIG is currently developing). • Provide a tool that enables the entry of study details (e.g., human/animal model, chemicals to be used) and indicate the required forms that must be completed and submitted. • Provide automated form-completion tool sets to collect and manage all necessary forms for approval within the institution (e.g., the IRB), with collaborating institutions, and with the funding agency and other external agencies (e.g., the Food and Drug Administration). • Maintain the forms in a controlled repository, providing easy, secure access to all regulators. If any changes occur (e.g., a protocol is updated, an office moves), the changes would be published to all stakeholders, as required. • Manage collaborative review and approval of regulatory requests. Maintain the library of forms for ready access by all stakeholders.

4. Manage Financial Reporting

Provide a set of support tools that are integrated with study management tools (e.g., protocols, schedules, supply lists) to remove as much of the management burden from researchers as possible. This may include support for establishing contracts for supplies, managing reporting to grantors, and managing time reporting.

- Support the automated establishment of study fund account structure requests and budgets, based on the study design and protocol. Include supplies and labor hour plans.
- Provide automated tools to capture expenses for supplies, facilities, and other items and transmit them to standard accounting systems. Include documentation and applications needed for specific requests.
- Provide work documentation tools to support contract and grant reporting requirements.
- Capture invoice details and support generation of payment requests.
- Provide information required to close accounts at the conclusion of the study.
- Facilitate administrative support, including monitoring the projects using financial tools, providing on-request reports of current financial status, prompting the ordering of needed supplies (i.e., reagents), and documenting the electronic receipt of invoices and payments.

5. Request, Gather, or Design Samples

Provide integrated environments for collection, design, and management of samples, from the initial definition of need through processing and long-term storage or disposal.

- Provide automated tools to develop tissue management plans, including accession, processing, numbering, scheduling, tracking, analysis requirements, and supply and equipment needs.
- Provide automated support to establish strain requirements, identify strain resources, make strain requests, and document their receipt and use.
- Provide automated support for identifying organisms needed, resources for obtaining the organisms, and the processes, tools, equipment, and facilities needed to manage the organisms and document their care.
- Provide a mechanism to track work performed at another facility as well as a way to store data returned from the other facility.
- Provide automated support for requesting and managing synthetic DNA sequences and associated data, which would be provided in an interoperable format that could be automatically added to a caBIG-Plus–compatible library.
- Manage the collection and integration of previously collected genomic and proteomic data, including all approvals for use and provenance of data sets.
- Support the design of mutants by providing easy access to design tools. Provide a standardized way to store and use mutant data so that results can be readily compared across laboratories.
- Manage the request, delivery, use, and limit of potentially hazardous materials to avoid holding more materials than is permitted by regulations.
- Ease collection of external resources. In addition to automating the discovery of supplemental data, services could be provided that enable, among other things, the submission and automated request of synthetic DNA and the ordering of whole organisms or animal models to be used in experimentation.

6. Conduct Research and Perform Experiments

Provide support tools for researchers as research is performed. Integrate the tools with the rest of the caBIG-Plus environment to eliminate duplication of data entry and unnecessary data conversion or integration activities.

- Retain an automated material repository, alerting lab management staff when a resource is low (i.e., a specific reagent) and providing an automated capability to order replacement supplies.
- Ensure that all data are collected following standards.
- Manage the collection of heterogeneous data sets from disparate devices, annotated files, and direct input from multiple sources.
- Provide a caBIG-Plus–compatible data set infrastructure so that data are already in exchangeable form as they are added to the repository.
- Provide electronic notebook capabilities, complete with configurable workflow, data provenance, and sharable work pages.
- Provide tools to support the management of tissue samples, including aliquot numbering, freezer inventories, capture of sample descriptions, and interpretation of results. Use standard data structures to capture and integrate data (e.g., genomic and proteomic data), expanding on the current caTISSUE suite of tools.
- Provide interfaces to standard commercial lab systems (along with the necessary documentation of patient consent).
- Provide tools for de-identifying and anonymizing data if required.
- Catalog and capture all images (e.g., scans) used in the study and record appropriate metadata (i.e., source, equipment used, dates).
- Provide a schedule for specific activities (e.g., performing interventions, capturing observations) that can be tailored for use by the entire team or by individual members of the team.
- Manage the environmental requirements for animals used in the study, including food, schedules, interventions, and environment.
- Provide tools for capturing, coding, and analyzing adverse events. Provide tools to analyze data to seek patterns in lab or clinical findings that might indicate adverse events. Use automated forms to provide data to appropriate researchers, clinicians, and regulators if adverse events are suspected or detected. Maintain a searchable catalog of all suspected or actual events and link these to specific data about protocols in use, site of the study, and clinical descriptions.
- Manage participant recruitment, accession, randomization, and retention. Manage data collection from participants (e.g., by using diaries). Manage interactions with participants' clinicians and principle researchers.
- Provide a set of tools for data validation and error correction that can be used as close to the data collection source as possible (e.g., the bench or the site of a clinical intervention). Track all data problems, providing regular reports to researchers and managers. Document provenance of all data for later analysis and regulatory reporting.

7. Conduct Analysis
Provide an integrated analytical environment with well-documented tools that can be accessed in a secure grid.
• Provide pathway analysis tools that can be pulled down from standardized libraries and connected to instrumented data sets and that output standardized results that can be plugged into a caBIG-Plus–compatible data structure. Implement these tools in a workflow structure for ease of use by researchers and clinicians. • Manage the identification, collection, and integration of data (e.g., normalized values for comparisons) from public repositories to support the study. • Capture analytical results from tissue data (in de-identified form, if necessary). Associate all samples with the appropriate study step, source, and analysis technique used. Collect all tissue data into a structured caBIG-Plus–compatible structure that facilitates analysis by a standardized tools library, sharable by the entire study team. • Provide tools to support the characterization of molecular targets, including tools that capture data sets, compare data sets with data sets from previous studies, and import data from analytical devices and equipment. • Provide standardized access to genomic analysis tools, structuring the inputs and outputs so that the tools can be used in a seamless manner to provide interoperable data sets. • Provide a way to acquire proteomic data; capture gel electrophoresis results, microarray data, and other data; provide libraries to manage, share, and analyze large data sets and evaluate collected images (e.g., capturing annotations); and provide clear data provenance throughout the process. • Manage a library of statistical and biomedical data analysis tools that can be used in a standardized form throughout the analysis process, while maintaining data provenance. Manage libraries of interim results and enable development of sharable workflow components to enable replicable analytical processes across teams.
8. Publish Results
Provide tools to manage the workflow and consolidate study results for publication.
• Support the collection of data for the study manuscript, providing analytical and formatting tools. • Maintain a repository of templates for both data and manuscript submission, with links to submission locations. • Establish a collaborative environment for developing and reviewing the study manuscript, while controlling access to preliminary results, capturing review comments, maintaining a workflow for circulating drafts, and providing access to study data sets for verification of results. • Manage the publication of the study manuscript to controlled repositories, which will be opened for broader use when the study is published. • Provide workflow tools for submitting the study manuscript to journals for publication, managing the review process, and responding to reviewers. • Assist in the publication of data and results. Providing a repository of the requirements and templates for submission of study manuscripts to various journals, and providing links for submitting data to the caBIG-Plus repository in the appropriate format, would speed the accessibility of information to the biomedical community.

The implementation of these processes would greatly improve researchers' ability to collaborate with one another, move through the regulatory processes, acquire and manage data, and complete their research in a timely manner. The next section provides scenarios to illustrate the difference between current processes and caBIG-Plus processes, as envisioned by this conceptual view.

3. "Day in the Life" Scenarios

The following scenarios illustrate how various caBIG stakeholders and users could accomplish research tasks in the caBIG-Plus operational environment. The six scenarios are as follows:

1. An executive from an academic medical center is financing a suitable research infrastructure.
2. A network administrator is monitoring his or her institution's interactions with caBIG-Plus.
3. A software developer from a leading university is connecting to caBIG-Plus.
4. A biomedical researcher is accessing phenotype data.
5. A clinical researcher is identifying potential participants for a clinical trial.
6. A vendor executive is deciding whether to upgrade a product to enable integration with caBIG-Plus.

3.1 Executive from Academic Medical Center: Financing a Suitable Research Infrastructure

Table 3-1. Scenario 1: Academic Medical Center Executive Financing a Suitable Research Infrastructure

Step	Description
1	Technical staff from several university clinical and research departments request approval to establish a research infrastructure that can be connected to caBIG-Plus.
2	Department managers provide a research plan, showing how participation in caBIG-Plus will improve access to data and collaborative research resources.
3	The dean reallocates resources to provide (1) access to legacy applications (as appropriate), (2) servers for interacting with caBIG-Plus, and (3) security technicians to ensure that appropriate security measures are put into place.
4	The research community leverages the caBIG-Plus environment to perform collaborative research that would not have been possible without access to caBIG-Plus data and tools. The university is able to attract more research support and leverage technical support.

Without caBIG-Plus	With caBIG-Plus
Licensed information technology products, which can be extremely expensive, are required for research labs. These products are not designed to a common interoperability specification; therefore, additional resources often are required to make them work with other tools in the research environment.	caBIG-Plus provides basic tools to the university research community that can be customized as needed for the university's environment. Additional licensed tools that are certified as caBIG-Plus compatible can be added for specific purposes and readily integrated without additional expense. Vendors also may provide caBIG-Plus–compatible interfaces, enabling their tools to be used in the caBIG-Plus environment.

3.2 Network Administrator: Monitoring caBIG-Plus Interactions

Table 3-2. Scenario 2: Network Administrator Monitoring caBIG-Plus Interactions

Step	Description
1	The network administrator accesses an on-demand security report that enumerates pending requests for access to resources (data and services) provided by his or her institution.
2	The administrator updates the access privileges of users at the institution (this includes the deletion of accounts for users no longer at the institution). The administrator also peruses the security procedures of external institutions to determine the maximum level of access that will be afforded to users at his or her institution.
3	The administrator quickly and easily retrieves a report summarizing the extent to which institution users are able to access remote data and services efficiently.
4	Based on this report, the administrator uses a grid configuration tool to establish policies that govern how best to answer internal requests for external resources. Before deploying these policies, the administrator evaluates their impact by running simulations.
5	The administrator generates a report that indicates which service providers seem to be most (and least) reliable. The administrator shares the report with the institution's research community through online collaborative tools and processes to help other administrators configure their systems, quickly identify and resolve problems, and select reliable service providers.

Without caBIG-Plus	With caBIG-Plus
• Each institution maintains its own security database. Researchers who want to access external data must either establish an account with the external institution or receive a copy of the data, whose access cannot be controlled by the data producer. • To the extent that external services are available, the policies that govern which service providers to access, and in what order, are hard-coded into client tools without considering the configurations of these providers. • Service providers with reliability and performance problems are often ignored completely. When their services are used, the feedback they receive is neither timely nor publicly reviewable.	• Each institution authenticates its own users; therefore, each user needs only a single account. However, each institution establishes its own authorization policy that governs who can access its data and services. • Monitoring tools track the reliability and performance of service providers. Reliability and performance data are then fed into a grid configuration tool so that the administrator can easily establish, test, and deploy configuration policies. • Reliability and performance feedback is shared with the research community in a public forum, which encourages the rapid resolution of any problems that arise and the selection of reliable service providers.

3.3 Software Developer: Connecting to caBIG-Plus

Table 3-3. Scenario 3: Software Developer Connecting to caBIG-Plus

Step	Description
1	The software developer from a leading university wants to share a new and useful service (e.g., a technique, an algorithm) for processing data.
2	After becoming aware of caBIG-Plus, the developer downloads the caBIG-Plus software development kit (SDK) from a public Web site. All SDK components are installed on the developer's computer.
3	Using the Enterprise Vocabulary Services (EVS), the developer determines which data elements will serve as input and output parameters for the new service.
4	The developer uses the SDK to implement and test the service. Because the developer is now a member of the caBIG-Plus community, his or her tests are based on real data (cleared for public release).
5	Satisfied with the service, the developer publishes it to the university research community's service registry along with the required metadata to help users discover the service via ad hoc search or automated discovery mechanisms.

Without caBIG-Plus	With caBIG-Plus
• The new service is published in a peer-reviewed journal or conference proceedings. The service is almost always tied to a specific operating system and programming language. • Assuming the developer publishes the service as a Web service, the input and output formats are unique to the service. As a result, each user of the service must transform his or her data into the required format before invoking the service. Similarly, the results must be transformed before they can be handed off to another service.	• The developer can easily advertise the existence of new services. These services can be accessed regardless of operating systems and programming languages used by other caBIG-Plus participants. • Because all caBIG-Plus components are provided in an all-in-one SDK, installing appropriate software to connect to caBIG-Plus is straightforward and does not require extensive training. This approach reduces the variety of software versions in use and increases the stability and consistency of the caBIG-Plus environment. • Because services' input and output parameters reference common data elements, it is easier to find relevant services and invoke them using one's own data. Similarly, individual services can be combined more easily to produce a complex workflow. • The researcher has fewer concerns about interoperability.

3.4 Biomedical Researcher: Accessing Phenotype Data

Table 3-4. Scenario 4: Biomedical Researcher Accessing Phenotype Data

Step	Description
1	Using a caBIG-Plus discovery method, the biomedical researcher identifies potential cases from anonymized data repositories across caBIG-Plus.
2	The researcher accesses the caBIG-Plus centralized IRB system and describes his or her proposed research on an online, standardized form. The researcher submits the form electronically to the IRB. The researcher can check the status of his or her research request online at any time.
3	The IRB accesses, reviews, and approves the research, thereby providing the researcher with access to the full set of requested anonymized data.
4	The researcher downloads data from multiple institutions into a standardized, normalized data set that contains metadata indicating the data provenance and provides key values (e.g., normal ranges).

Without caBIG-Plus	With caBIG-Plus
• The biomedical researcher must phone or email colleagues to identify potential data sources and negotiate with data owners and individual IRBs to obtain data. • The researcher obtains data in multiple, inconsistent formats and spends many days normalizing the data and understanding the methods used to collect them (e.g., identifying the normal range of lab readings for the equipment used).	• The biomedical researcher can quickly and easily identify existing data that supplements data from an ongoing study, saving time and money for the researcher and enabling conclusions to be drawn at a higher power. • caBIG-Plus enables the researcher to identify and download data efficiently in a standardized format. The researcher can quickly identify the best data sets for his or her needs, as annotated in caBIG-Plus collaborative workspaces, and rapidly proceed through research steps. Receiving all data in the same format enables the researcher to perform accurate analyses without misinterpreting data from other groups. Because the researcher is able to work more efficiently, he or she needs less technical support and fewer lab technicians. He or she also is able to conduct more research, more productively. Because most of the "hassle factor" is eliminated, the researcher finds a research career more attractive. • The IRB automates approval processes, enabling approval documents to flow quickly, smoothly, and correctly through the process. The researcher is pleased because there is minimal disruption to his or her research, while complying with all requirements levied by funding authorities. IRB members are pleased because the standardization of the format and workflow ensures that protocols are in the required format, all required data and signatures are provided, all regulatory requirements are met, and the research request is readily comparable with previous requests.

The centralized IRB concept will require substantial work with regulatory and institutional bodies. caBIG-Plus could provide the environment necessary for implementing the agreements reached by necessary stakeholders.

3.5 Clinical Researcher: Identifying Potential Participants for a Clinical Trial

Table 3-5. Scenario 5: Clinical Researcher Identifying Potential Participants for a Clinical Trial

Step	Description
1	By accessing de-identified medical record summaries on caBIG-Plus, the clinical researcher identifies a cohort of patients who might be eligible for the clinical trial.
2	Using the honest broker service, the clinical researcher contacts a patient's physician to determine whether the patient is eligible to participate in the clinical trial. The physician already posted a metadata profile that indicates whether he or she is interested in participating in clinical trials, the types of trials he or she is interested in, and his or her current participation in clinical trials. This enables the clinical researcher to leverage the physician's trained staff (e.g., if a physician's staff is already trained to support oncology clinical trials, it is easier for the physician to participate in a new oncology clinical trial).
3	The physician evaluates the proposed clinical trial and the patient selected for evaluation. The physician validates the patient's eligibility and contacts the patient to determine his or her level of interest. The complete protocol and set of disclosure forms is available online for the patient to review.
4	The physician contacts the clinical researcher to indicate the patient's agreement to participate in the clinical trial.
5	The clinical researcher obtains access to the patient's electronic health records and additional data for automatic enrollment and eligibility verification, as required in the approved protocol.

Without caBIG-Plus	With caBIG-Plus
The clinical researcher must use email, snail mail, Web sites, and meetings to communicate with physicians to find patients who are eligible to participate in a clinical trial. This process is slow, expensive, and reaches many patients who are not eligible to participate in the trial and omits many patients who are eligible and would participate if they were aware of the trial.	The clinical researcher can rapidly target a specific group of patients who might be interested in participating in the clinical trial. Patients can quickly access study protocols and respond to requests for participation, and researchers can access electronic health records to investigate correlations in study data. The result is the recruitment of patients who are interested in participating, and eligible for participation, in the clinical trial. The facilitation of the recruitment process improves patient support for the trial and increases patient retention because there are fewer false starts and faster determinations of eligibility.

3.6 Vendor Executive: Upgrading a Product for caBIG-Plus Integration

Table 3-6. Scenario 6: Vendor Executive Upgrading Product for caBIG-Plus Integration

Step	Description
1	The executive of a company specializing in software and tools for the biomedical community is faced with the decision of whether to upgrade an existing product to enable integration with caBIG-Plus.
2	The company conducts a return-on-investment (ROI) analysis to demonstrate to the vendor executive the benefits of integrating the product with caBIG-Plus rather than providing a proprietary interface. The ROI indicates that product integration with caBIG-Plus will increase sales because caBIG-Plus users will be able to find and integrate the product easily into their workflows. The ROI also indicates that caBIG-Plus integration will enable the vendor to offer the product on a subscription basis through caBIG-Plus, thus increasing market share.
3	The vendor executive supports product integration with caBIG-Plus.
4	The vendor is able to market a product that can be incorporated into the workflows of caBIG-Plus users and interoperate with other tools throughout the caBIG-Plus environment.

Without caBIG-Plus	With caBIG-Plus
• The vendor must develop point-to-point interfaces with numerous databases and other vendor tools. • Individuals are reluctant to invest in tools that require expensive integration into their work environments, especially if there are concerns about a lack of standardized data definitions and structures.	• Scarce development resources can be applied to enhancing the product, instead of developing multiple interfaces. • Additional market opportunities are provided by the new interoperability. The more the tool is incorporated into the workflows of the biomedical community, the more likely the vendor is to make additional sales.

These scenarios illustrate the transformative power of a caBIG-Plus environment and how such a solution implemented across the entire biomedical research community could act as a "matchmaker" among physicians, researchers, information technology (IT) professionals, and patients. The scenarios demonstrate how easily accessed and easily installed tools enable research centers to use their funds and IT staff most effectively. Researchers can quickly access the best data sets for their needs and can access appropriate colleagues with whom to share knowledge. Researchers can manage regulatory approval and other administrative processes online using automated workflow, helping reduce the hassle factor and enabling more and better research. Technical staff can easily monitor their networks and manage performance levels. Clinical researchers and patients alike can quickly "match up" for a study. Finally, vendors can leverage the caBIG-Plus environment to achieve their objectives of efficiently providing interoperable solutions to the entire biomedical research community.

4. Critical or Limiting Factors: Assumptions and Risks

This section highlights our assumptions about and identifies the high-priority risks associated with expanding caBIG. If these assumptions do not hold, or if the risks materialize, progress in achieving the caBIG-Plus vision will be considerably hampered.

4.1 Assumptions

The following assumptions are necessary for the realization of caBIG-Plus:

- **Funding.** Funding must be available to create caBIG-Plus.
- **Open Source Technology.** This technology can provide the basis for caBIG-Plus development. Appropriate levels of investment in open source technologies must be made, key open source technologists must be willing to scale the steep learning curve to contribute to caBIG-Plus, and a highly respected technical leadership must evolve to vet contributions, encourage involvement by gifted developers, and work with the research community. To support open source technologies effectively, NCRR should continue to provide contractual support to encourage open source development, maintenance, and enhancement. In addition, NCRR should fully leverage the new Clinical and Translational Science Awards to encourage the development of open source research tool interoperability. Fostering interoperability across open source tool sets and with vendors is critical. The intent is not to replace commercial tools but to ensure that both open source and commercial tools are interoperable.
- **Common Vocabulary.** Common vocabulary must be developed so that biomedical researchers and healthcare providers can access the same data sets using common terminology. This assumes that data are organized and labeled in a consistent manner. First, well-defined standards are needed for common data types (e.g., the Minimum Information About a Microarray Experiment [MIAME] standard for microarray experiments). Second, each domain must establish a common vocabulary and thesaurus; the thesaurus could be developed in a modular manner, similar to the approach used to develop the NCI Thesaurus. Third, within each domain, data brokers must be established so that researchers and healthcare providers do not need to interact with every data producer. Finally, rough correspondences need to be established across domains; these loose connections can be formalized when the need arises. Harmonization of vocabularies is necessary both within and between domains. Although there are many similarities in terms (e.g., the definition of "white blood cell" across domains), it is important to ensure that definitions are the same across domains (e.g., pathology versus hematology) before identifying them as synonyms in the terminology tool set. Automated systems within a domain do not necessarily define terms in the same manner; tool providers could be encouraged to map their terminologies to the standards to improve interoperability across the board.
- **Governance.** Governance is implemented by recruiting capable leaders and keeping them involved in the governance process. Resolving the differences among varying governance approaches, establishing governance policies, locating additional funding, managing system security, and attracting competent open source developers will be challenging.

All these assumptions present significant challenges, especially developing a stable and sufficient source of funding. If stakeholders are aware of the benefits of caBIG-Plus though, funding should be easier to obtain.

4.2 Risks

With any endeavor as large and complex as the expansion of caBIG, the risks associated with achieving success are significant and must be monitored and mitigated aggressively. Table 4-1 presents the highest priority risks, along with recommended strategies to prevent the risks from materializing.

Table 4-1. caBIG-Plus Risks and Recommended Mitigation Strategies

Risk	Description and Mitigation Recommendations
Common Vocabulary	If vocabularies cannot be mapped or shared (i.e., if differences in terminology, semantics, and underlying data models are not resolved), data from different sources cannot be used effectively. The lack of data interoperability would force each user to replicate the work of other users in order to understand each data source and harmonize the data.
	Recommendations for Mitigation: • Establish a standard methodology for vocabulary development and provide a set of tools to support this development process. Maintain a Vocabulary Advisory Group to recommend methodology and tools and review the resulting vocabularies. • Use description logic (e.g., Web Ontology Language [OWL]-DL) as the underlying language for ease of extensibility. • Focus first on data type–specific vocabularies. Then build domain-specific vocabularies and link these vocabularies to existing systems. • Coordinate efforts with the open biomedical ontologies community.
Security	Without strong security measures, caBIG-Plus will not be trusted or used. Likewise, use will be discouraged if a security solution makes it difficult for users to access or manage caBIG-Plus data.
	Recommendations for Mitigation: • Form a Security Advisory Committee. The committee should be composed of external experts who play significant roles in non-NIH infrastructures (e.g., Department of Defense and stock exchange infrastructures). Leading members of key technology standards groups also should participate. The Security Advisory Committee would provide access to best practices and new technologies and review and approve security plans. Existence of such a group could provide a forum for discussing policy issues and building consensus on approaches for managing security across caBIG-Plus. Ultimately, participating institutions will have to adopt the technologies and policies that are appropriate for the caBIG-Plus environment. • Establish a strong governance structure to address the difficulty of establishing and enforcing security policies across such a disparate community. Capstone policies and enabling standards, guidelines, and procedures should be developed along with identity and access management mechanisms. Certification and accreditation guidance for credential providers, and the associated assurance level of the credentials themselves, are examples of key governance policies in the security domain. • Establish robust security monitoring and reporting processes and procedures to identify and respond to system attacks. • Allocate the right mix of resources. For example, implementing the identity federation necessary for authorization and authentication is complex. It involves the collaboration of individual institutions. It also requires that caBIG-Plus identify an implementation team that has a clear vision of identity federation and that understands business use cases, existing infrastructure limitations, security requirements, regulatory compliance, legal implications, and the hands-on skills needed to create the federated architecture.[13] • Consider the maturity of technologies. Few security technologies, as discussed in the security technology evaluation white paper, have been deployed in a large-scale production environment. Many evaluated technologies have limited development resources, which would impact the quality and capability of production support once the software is released.[14]

Risk	Description and Mitigation Recommendations
Usability and Ease of Installation	If the user interface is not easy to use or if installing caBIG-Plus is too difficult and time-consuming, potential users may not participate. caBIG-Plus can provide a forum for users to explain their interface needs and comment on all types of caBIG-Plus–enabled tools (e.g., commercial tools, open source tools) and their usability in the overall workflow.
	Recommendations for Mitigation: • Engage users with a wide range of expertise to analyze multiple use cases to ensure that solutions and tools are designed with a robust understanding of end user requirements. • Create a second tier of adopters to represent end users and provide feedback on the ease of installation and use of the tool suite. • Invest in the development of a user-friendly graphical user interface to encourage use by scientists, students, and other users. • Focus on the development of a few robust, hardened tools and build from there (as opposed to continuously upgrading many tools). • Simplify the installation process in the following ways: – Package caBIG-Plus into a single installation bundle, including all applications on which caBIG-Plus depends (e.g., Web servers, databases). Provide an overview of how to configure these applications. If the environment includes commercial tools, provide a way to bundle them (e.g., through agreements with vendors). – Increase caBIG-Plus support resources (e.g., comprehensive frequently asked questions, animated Web tutorials, personnel). Limit reading materials to installation troubleshooting, and supplement these materials with animated tutorials that contain actual data. Develop more Web user interfaces to reduce the amount of command prompt instructions and to guide novice users through installation and operations. – Provide a user-friendly environment for caBIG-Plus users to communicate with one another in real time to find solutions for installation and usability problems. – Modify caBIG-Plus software so that fewer non-critical fields are required in the Unified Modeling Language (UML) model. By providing model submissions with less descriptive content, it is less likely that users will abandon the installation.
Scalability and Technical Performance	If users experience significant performance shortfalls when attempting to move large data sets (e.g., images) through the XML-based infrastructure or when trying to use real-time Web services, caBIG-Plus adoption rates will slow.[15]
	Recommendations for Mitigation: • Compress XML messages when resources (e.g., disk space, network bandwidth) are scarce. Evaluate available compression techniques[16] to determine which techniques caBIG-Plus users are expected to support. • Evaluate new tools for compression quality and availability (even if they are not open source) to speed the processing of XML messages. Mandate the approaches caBIG-Plus users must support.

Risk	Description and Mitigation Recommendations
Non-Use and Institutional Resistance by the Non-Cancer Community	Potential caBIG-Plus users may resist adopting the tools and processes within the caBIG-Plus environment because they are concerned that data will be improperly used, the benefits are unclear, and a compelling case to use caBIG-Plus has not been made. If a highly effective transition plan, endorsed by senior-level stakeholders, is not developed and executed adaptively, adoption will be slow and benefits will be realized later than anticipated.
	Recommendations for Mitigation: • Identify major success stories (e.g., disease biomarker, prognostic predictor, cancer subclass discovery) from caBIG-Plus users (e.g., bench scientists) to identify best practices associated with increasing caBIG-Plus adoption and use in similar environments. • Develop and highlight a "killer application" that is compelling enough to dramatically increase caBIG-Plus use, publicize caBIG-Plus capabilities to new communities, and demonstrate to potential users the benefits of involvement in the caBIG-Plus community. Establish a robust coordination process for identifying high-value tools and data among multiple communities. • Mandate the use of appropriate caBIG-Plus components as a condition of accepting grant funding. • Collect data sets from all public repositories so that caBIG-Plus becomes known and is used as a "one-stop shop" for data and therefore achieves the critical mass of exchangeable data necessary for drawing more users. • Promote the use of caBIG-Plus in teaching environments to get students acclimated to, and comfortable with, caBIG-Plus and establish it as the standard for a new generation of researchers. • Develop a stakeholder management plan and a robust communications plan to increase caBIG-Plus awareness, highlight the benefits of participating in caBIG-Plus, highlight available tools and data, and collect feedback and suggestions from all users. • Address data sharing concerns in the following ways: – Allow data providers to publish a subset or summary of their data and allow users to obtain the entire data set only after establishing an agreement with the data provider. – Introduce a file versioning control system to track data sets that are derived by extracting a subset of the data from another data source or by combining data from multiple sources so that original data providers are properly attributed. – Require users to notify data providers when their data are analyzed or interpreted to enable data providers to validate the results. – Implement a copyright procedure similar to that of Creative Commons[17] to allow data providers to specify permissions. – Allow data providers to withhold data prior to publication or provisional patent filing.

Addressing all of these risks will require a significant investment. Failure to do so will derail the success of caBIG-Plus. It is imperative to quickly establish robust risk management processes and senior-level accountability to ensure that risks are thoroughly addressed and aggressively managed throughout the caBIG-Plus implementation.

5. Conclusion

caBIG has benefited the cancer community by enabling collaboration in the community and by speeding the dissemination of novel discoveries through data exchange and development of data analysis tools.

The non-cancer research community faces many of the same issues the cancer research community faces. caBIG expansion into caBIG-Plus will benefit the entire biomedical research community. Rapid advances may result as researchers join a more collaborative community, thus reducing work repetition, sharing work flows, and identifying useful results. Researchers previously unable to leverage computational capabilities will have the opportunity to do so through a more user-friendly environment.

To ensure caBIG-Plus success, it is critical to make the investments necessary to fully address critical success factors (e.g., security) identified by the caBIG community. In addition, some processes that are currently defined more narrowly for the cancer research community will have to be expanded to support the needs (e.g., terminology) of the larger research community. Creation of a usable common data exchange and data analysis platform will be supported by researchers across domains (e.g., the pharmaceutical industry, clinicians).

If caBIG-Plus is not accepted across all biomedical research domains, research will continue to be conducted as it is now, using silos of data that cannot be shared readily across laboratories or across the translational continuum. Collaboration and the ability of researchers to find other researchers with similar interests will remain limited. Eventually, the realization will be made that a platform offering many of the same capabilities as caBIG-Plus is needed and will be designed at that time. Resistance to caBIG-Plus acceptance may be even greater by then because researchers and institutions will have invested more time and money into developing their legacy systems.

At some point, the capabilities offered by caBIG-Plus, as described in this report, will need to be established. The cost of doing so now is significant; however, the alternative is even more costly—retroactively implementing a broad-based biomedical informatics grid when the biomedical community reaches a point of desperation. As Crowley and colleagues point out:

> "Only timely, integrated and system-wide investments can deliver the tacit promises of improved health care to our nation that attend our rapid basic science advances; marginal investments in the already poorly functioning and overloaded system will not."[18]

Appendix A. Contributors

This report was prepared by a team of MITRE staff members. Team members are as follows:

- Amy Aukema
- Brandon Higgs, Ph.D.
- Robert Mikula, M.S.
- Peter Mork, Ph.D.
- Olivia Peters, M.S., Project Lead
- Jean Stanford, Project Manager
- Marion Warwick, M.D.

The MITRE team wishes to acknowledge the technical review provided by Don Faatz, a MITRE technical staff member.

Appendix B. Acronyms

caBIG™	Cancer Biomedical Informatics Grid™
caTISSUE	Cancer Tissue Database
EVS	Enterprise Vocabulary Services
FFRDC	Federally Funded Research and Development Center
IRB	Institutional Review Board
IT	Information Technology
MIAME	Minimum Information About a Microarray Experiment
NCI	National Cancer Institute
NCRR	National Center for Research Resources
NIH	National Institutes of Health
OWL	Web Ontology Language
ROI	Return on Investment
SDK	Software Development Kit
SOA	Service Oriented Architecture
UML	Unified Modeling Language
XML	Extensible Markup Language

Appendix C. Key Concepts

Some of the emerging technologies incorporated by the Cancer Biomedical Informatics Grid™ (caBIG™) are as follows:

- **Open source**, which is a method of designing, developing, and sharing computer software in a manner that is open to the public and available without charge. Open source systems generally evolve through community cooperation, and caBIG is modeled on this approach. Many major open source software products rely both on volunteer developers, who work on their own time, and paid contributors, who work on the software as a part of their company's strategic plan (e.g., IBM contributes a large amount of open source software to the Linux open source operating system because it fits into its overall technical strategy[19]). Open source systems generally publish their source code (the set of computer instructions that can be compiled and integrated into a working system) on the Internet under a licensing agreement that usually is without charge. Many open source communities provide extensive software training, documentation, and maintenance and enhancement capabilities. Examples of open source products are Linux (a major operating system), Mozilla (an Internet browser), and Mediawiki (which is used to run tools such as Wikipedia).
- **Service oriented architecture (SOA)**, which is a way of decomposing information technology (IT) applications into sets of interoperable services. For example, a Web site could provide a currency conversion service by linking to a service at a bank that provides conversions on demand. By using the bank's service, the Web site owner would not have to know anything about currency rates in today's market. Instead, the owner could take advantage of a published application programming interface (API) provided by the bank, submit the dollar amount to be converted and the currency of interest (e.g., the euro), and receive an answer, which is then provided to Web site users. Service offerers would publish a set of metadata[20] in a registry that indicates the type of services they offer, where the services are located on the Web, how to call them, and the degree of accuracy or precision that can be expected from them. Systems such as GoogleEarth and Amazon.com offer many services over the Web. It is not necessary to use the Web to provide services, but many organizations do. Typically, a service encapsulates a business process, which may be very atomic (e.g., currency conversion) or more complex (e.g., arranging for payment via credit cards). caBIG services will be offered within a grid environment, as described in the next paragraph.
- **A grid**, which is a "hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities."[21] Computer systems on the grid use a common set of tools, such as common security and authentication, to exchange services and data. caBIG participants can put data sets on the grid (called "exposing" the data), which then can be accessed via services, as described previously, by caBIG authorized users. Participants also can share access to computational resources (e.g., computer systems with available computing cycles) for performing large computations (e.g., protein folding calculations). The use of a grid essentially enables the creation of a virtual enterprise, in which all grid participants appear visible to one another but not to the outside world.

Appendix D. Endnotes

[1] Zerhouni EA, "Translational and Clinical Science—Time for a New Vision," *New England Journal of Medicine* 353(15):1621–1623, October 13, 2005.

[2] Crowley WF Jr, Sherwood L, Salber P, Scheinberg D, Slavkin H, Tilson H, Reece EA, Catanese V, Johnson SB, Dobs A, Genel M, Korn A, Reame N, Bonow R, Grebb J, Rimoin D, "Clinical Research in the United States at a Crossroads: Proposal for a Novel Public-Private Partnership to Establish a National Clinical Research Enterprise," *JAMA: Journal of the American Medical Association* 291(9):1120–1126, March 3, 2004.

[3] See, for example, Zerhouni EA, op. cit.

[4] Culliton BJ, "Extracting Knowledge from Science: A Conversation with Elias Zerhouni," *Health Affairs* 25(3):w94–w103, 2006, http://content.healthaffairs.org/cgi/content/abstract/hlthaff.25.w94.

[5] For more information on caBIG, go to https://cabig.nci.nih.gov.

[6] The MITRE Corporation, *caBIG™ Overview*, May 2006, http://www.ncrr.nih.gov/CRinformatics/mitre.asp.

[7] Institute of Medicine of the National Academies, Kohn LT (editor), *Academic Health Centers: Leading Change in the 21st Century*, The National Academies Press, pp. 77–91, 2004.

[8] Zerhouni EA, "The NIH Roadmap," *Science* 302:64, October 3, 2003.

[9] Buneman P, Abiteboul S, Szalay A, Hagehulsmann A, "Laying the Ground: Semantics of Data," *Towards 2020 Science*, Microsoft Corporation, p. 15, 2006, http://research.microsoft.com/towards2020science/downloads.htm.

[10] Campbell EG, Weissman JS, Moy E, Blumenthal D, "Status of Clinical Research in Academic Health Centers: Views from the Research Leadership," *JAMA: Journal of the American Medical Association* 286(7):800–806, August 15, 2001.

[11] The MITRE Corporation, *caBIG™: Opportunities and Challenges for Use Beyond Cancer*, June 2006.

[12] For a description of caTISSUE, go to https://cabig.nci.nih.gov/workspaces/TBPT.

[13] caBIG™ Security Technology Evaluation White Paper, October 7, 2005, https://cabig.nci.nih.gov/workspaces/Architecture/Security_Tech_Eval_White_Paper_Provisional.

[14] Ibid.

[15] This problem stems primarily from two sources. First, all data are represented textually, even when the data are quantitative or temporal—data types for which more efficient representations are possible. Second, all data are marked with a starting tag and an equivalent ending tag. These tags make the data easily parsable by both humans and machines; however, they create redundancy.

[16] Cokus M, Winkowski D, *XML Sizing and Compression Study for Military Wireless Data*, XML Conference and Exposition 2002, December 8–13, 2002, http://www.idealliance.org/papers/xml02/dx_xml02/papers/06-02-04/06-02-04.html.

[17] Creative Commons, http://creativecommons.org/license.

[18] Crowley, et al., op. cit., p. 1125.

[19] "The IBM Linux Technology Center is a worldwide organization with teams in approximately 40 locations. It comprises some 600 engineers worldwide, of whom more than 300 work full-time on Linux as part of the open source community. The investment to expand Linux development in Brazil complements ongoing work at IBM's Linux Integration Centers, Linux Innovation Centers, and Linux Competency Centers, all of which help customers port applications to Linux." From "IBM Expands Linux Technology Center in Brazil," *The IBM LinuxLine*, May 24, 2006, http://www.dbta.com/linuxline/archives/5-24-06.html.

[20] "Metadata" is a term commonly used in the biomedical informatics field: "(1) Information about a data set which is provided by the data supplier or the generating algorithm and which provides a description of the content, format, and utility of the data set. Metadata provide criteria which may be used to select data for a particular scientific investigation. (2) Information describing a data set, including data user guide, descriptions of the data set in directories, and inventories, and any additional information required to define the relationships among these." See glossary at http://podaac.jpl.nasa.gov/glossary/.

[21] Wolfgang G, "DOT-COMing the GRID: Using Grids for Business," Sun Microsystems, Inc., http://www.sun.com/software/gridware/article.xml.