National Cancer Institute

caBIG™
cancer Biomedical Informatics Grid™

The caBIG™ Pilot Phase

Report: 2003-2007

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institutes
of Health

The caBIG™ Pilot Phase Report 2003-2007 was commissioned by the National Cancer Institute Center for Bioinformatics (NCICB) to report back to the many constituencies of the cancer community and to the nation on the progress made during the first three years of this pioneering endeavor. The Report is intended to set forth the goals and strategic direction of the caBIG™ initiative, to delineate its accomplishments and shortcomings, and to set the stage for broader adoption as it realizes the potential for transforming cancer research in coming years.

The Report is based on a review of caBIG™–related documents and presentations produced by NCICB in recent years; materials about caBIG™ produced by the caBIG™ general contractor and other external entities; and one-on-one discussions held during the spring and summer of 2007 with caBIG™ participants and observers from government, academe, and industry. The NCICB is grateful to the many individuals who shared their knowledge and insights about caBIG™ for the Report.

November 2007

caBIG™
cancer Biomedical
Informatics Grid ™

The caBIG™ Pilot Phase

Report: 2003-2007

Table of Contents

Executive Summary **4**
- caBIG™ Goals and Outcomes 4
- Central Tenets of caBIG™ 5
- Moving into Enterprise Phase 6

Chapter 1: Background and Rationale **7**
- The Inception of caBIG™ 7
- Cancer and the Shift to Molecular Medicine 8
- The Challenges of Connectivity 9
- The Vision and Mission of caBIG™ 10
- Overarching Objectives of caBIG™ 10
- The Principles of caBIG™ 10
- caBIG™ Philosophy and Culture 11
- Pilot Phase Goals and Summary of Outcomes 11
- Financial Investment 12

Chapter 2: Strategic Planning and Initiation **13**
- Strategic Approaches and Practices 13
- Key Participants 14
- Key Roles 15
- Assessment of Needs 15
- Launch Plan and Key Pilot Phase Deliverables 16
- Case Study: caBIG™ License 18

Chapter 3: Programs and Progress **20**
- Pilot Phase Stages 20
- Workspaces 21
- The Development Cycle 25
- Funding Mechanism 26
- Progress on Deliverables 26
- caBIG™ Connections to Other NCI Programs 27
- caBIG™ Connections to External Programs and Organizations 28
- Program Spotlight: Regulatory Data Exchange (RDE) Initiative 28
- Recognition of caBIG™ 30
- Program Spotlight: BRIDG 31
- caBIG™ in the Literature 32
- Case Study: The caBIG™ Community 33

Chapter 4: Hindsight **34**
- Program Management and Community Engagement 35
- Communications 35
- Business Processes 36
- Software Development 37
- Role of the Private/For-profit Sector 39
- Cultural Shifts 40

Chapter 5: Future Directions **41**
- The Launch of the Enterprise Phase 41
- caBIG™ Adoption Program 41
- Enterprise Support Network 41
- caBIG™ and the Patient Community 42
- caBIG™ as a Model 43
- Case Study: Early Adoption by Duke 44

Appendices

Appendix A – caBIG™ Tools and Technologies **46**
Appendix B – Organizations Participating in the caBIG™ Initiative **54**

Table of Figures

Figure 1. Molecular Medicine: Selected Domains Needed to Enable Translational Research **9**
Figure 2. Gaining Input from the Cancer Centers. **16**
Figure 3. caBIG™ Initial Plan **17**
Figure 4. caBIG™ Pilot Phase Key Activities **20**
Figure 5. caBIG™ Workspace Organization **22**
Figure 6. Representative Attendance on caBIG™ Workspace Teleconferences 2004 through 2006/2007 . . . . **23**
Figure 7. caBIG™ Software Development Flow of Activities **25**
Figure 8. caBIG™ Annual Meeting Attendance 2004 – 2007 . . . . **26**
Figure 9. caBIG™ Deliverables by Category **27**
Figure 10. caBIG™ Web site Visits **32**
Figure 11. Future Directions for caBIG™ **43**

EXECUTIVE SUMMARY

In 2003, the National Cancer Institute (NCI) conceived of the cancer Biomedical Informatics Grid (caBIG™) initiative as part of its mission to advance research on cancer and improve clinical outcomes for patients. The NCI recognized that the ability to connect people, organizations, and data through information technology would be critical to realizing the potential of Molecular (also known as "personalized") Medicine.

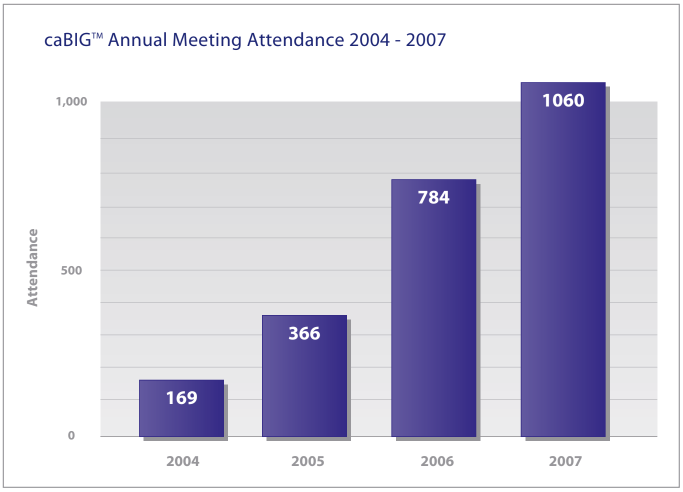
Over time, the need for a "worldwide web of cancer research," as caBIG™ has been described, has become increasingly urgent. Cancer continues to be a major problem globally, and the vast amount of data that researchers must collect, analyze, and store continues to expand as the pace of discovery accelerates.

The caBIG™ initiative has been an unprecedented effort. At the time of its launch, there was no such interconnected standardized biomedical informatics platform in place anywhere within the biomedical research community that could be used as an organizational or technical model. In that context, caBIG™ was—and continues to be—highly ambitious, especially in light of the cultural shifts within the research community that are required to make it a reality.

caBIG™ Goals and Outcomes

The caBIG™ initiative was formally launched in February 2004 as a three-year Pilot, overseen by the NCI Center for Bioinformatics (NCICB).[1] The objectives of the Pilot Phase were twofold: to test the ability of a complex informatics initiative to achieve measurable goals and deliverables toward enhancing cancer research, and to assess the opportunities and challenges of connecting a disparate biomedical community on a national and eventually international scale. Those objectives have been met, often beyond the expectations of the community, as follows:

- **Goal**: Illustrate that a spectrum of Cancer Centers with varying needs and capabilities can be joined in a common grid of shared data, applications, and technologies.
    - *Outcome: As of June 2007, there were over 190 organizations participating in the caBIG™ community (See Appendix B for complete listing). This community includes 51 Cancer Centers; federal agencies; and academic, not-for-profit, and industry entities, represented by close to 1,000 individuals.*



- **Goal**: Demonstrate that Cancer Centers, in collaboration with NCI, will develop new enabling software tools and systems to support multiple research organizations.
    - *Outcome: More than 300 software components have been delivered during the Pilot Phase, including over 40 end-user applications and a wide range of infrastructure components such as data standards and software development toolkits. Over 45 biomedical datasets have been delivered in caBIG™ compliant formats, derived from clinical and molecular studies, and are in use at several cancer research sites. These components were derived from over 5,000 analysis and requirement-gathering task orders issued to the community.*

      Due to the availability of this software at the close of the Pilot Phase, activities were under way in June 2007 to provide installation and support services

[1] *In 2007, NCI Center for Bioinformatics (NCICB) became part of the new overarching Center for Biomedical Informatics and Information Technology (CBIIT).*

for a wide range of adopting organizations, including a dedicated rollout of key tools, infrastructure, and interoperability framework to NCI Cancer Centers. The future availability of these tools and datasets over caGrid will enable an increasing number of investigators to share knowledge as it emerges.

- **Goal**: Demonstrate that Cancer Centers will actively use the grid and realize greater value in their cancer research endeavors by using this network to support powerful collaborations that are dependent on the sharing of data.
  - *Outcome: When caGrid (the data transmission network upon which caBIG™ works) was launched, several nodes (i.e., connection points where research organizations log onto the caBIG™ system) and software tools were available. As of June 2007, caGrid was extensively being used as a testing platform for the many caBIG™ software applications that will be grid-enabled to support the cancer research community in the Enterprise Phase.*
- **Goal**: Create an extensible infrastructure that will continue to be expanded and extended to members of the cancer research community beyond the NCI-designated Cancer Centers.
  - *Outcome: NCICB has actively collaborated with NCI programs (such as SPOREs,[2] The Cancer Genome Atlas,[3] the Cooperative Groups,[4] and Cancer Genetic Markers of Susceptibility[5]); other NIH Institutes and Centers (such as National Institute of Neurological Disorders and Stroke,[6] National Human Genome Research Institute,[7] National Heart, Lung and Blood Institute[8] and National Center for Research Resources[9]); other NIH-funded programs (such as Biomedical Informatics Research Network,[10] and Clinical and Translational Science Awards[11]); other research grid initiatives and other international informatics initiatives (including a formal agreement with the UK's National Cancer Research Institute[12]). Wherever possible, interoperability with such initiatives has been sought, and caBIG™ tools and infrastructure are consistently made available for either adoption or further development.*

Recognition of caBIG™

- *40+ Peer-Reviewed Publications[13]*
- *ComputerWorld Honors Program 2006[14]*
- *Three reports from The NIH National Center for Research Resources, 2006[15]*

Central Tenets of caBIG™

Interoperability is central to caBIG™; that is, compatibility among information technology tools used to collect, analyze, and share data. This compatibility provides a means of linking together all the scientists, clinicians, patients, and other participants so that they can conduct more dynamic, collaborative, and ultimately more successful research.

Among the hallmarks of the caBIG™ initiative has been the building of community. From initial outreach to Cancer Centers to identify the most pressing research needs, to the organization of community-based workspaces encompassing multiple disciplines to organize activities to address those needs, to the process for development and testing of software tools, caBIG™ has been *of* and *for* the cancer community.

[2] *http://spores.nci.nih.gov/.*

[3] *http://cancergenome.nih.gov/index.asp.*

[4] *http://www.cancer.gov/cancertopics/factsheet/NCI/clinical-trials-cooperative-group.*

[5] *http://cgems.cancer.gov/.*

[6] *http://www.ninds.nih.gov*

[7] *http://www.genome.gov*

[8] *http://www.nhlbi.nih.gov*

[9] *http://www.ncrr.nih.gov/clinical_research_resources/.*

[10] *http://www.nbirn.net/index_ie6.shtm.*

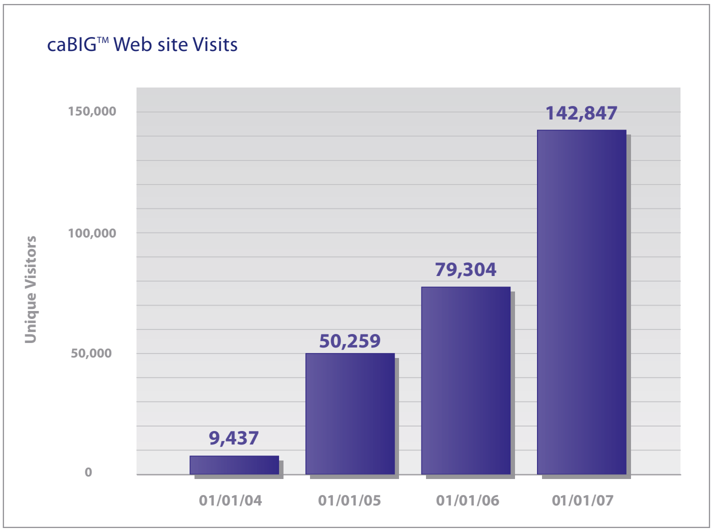
[11] *http://ctsaweb.org/about.html.*

[12] *http://www.ncri.org.uk/.*

[13] *http://caBIG.nci.nih.gov/Library/Library/caBIG_Scientific_Pubs.html.*

[14] *http://www.cwhonors.org*

[15] *http://www.ncrr.nih.gov/publications/informatics/caBIG.pdf; http://www.ncrr.nih.gov/publications/informatics/caBIG_OpportunitiesAndChallenges_12-26-06.pdf; http://www.ncrr.nih.gov/publications/informatics/caBIG-Plus_ConceptualView_12-26-06.pdf*



Another key characteristic has been the openness of the caBIG™ initiative, as reflected in the open source nature of the standards and software and the open access to the initiative for any constituencies within the biomedical community who wished to participate. The caBIG™ initiative was also marked by the dynamic nature of its management structure and operations, which retained a highly flexible capacity to change over time as conditions in the community changed, as well as to absorb lessons learned as the program evolved.

Moving into Enterprise Phase

The caBIG™ Pilot Phase concluded in March 2007, followed by a transition to an Enterprise Phase, which has built on caBIG™ accomplishments and lessons learned. In the Enterprise Phase, an expanding number of organizations—including additional Cancer Centers, the pharmaceutical and biotech community, and the commercial IT sector—are being invited to achieve connectivity via the broader adoption of caBIG™ tools, infrastructure, and interoperability framework. caBIG™ is also sharing its experience, expertise, and tools with the larger biomedical community to serve as a model, so that other disease-focused endeavors can advance more rapidly through a learning curve to build their informatics capabilities. It is likely that most of the tools and infrastructure of caBIG™ will be widely applicable beyond cancer.

In addressing the 2007 caBIG™ Annual Meeting, Dr. Elias Zerhouni, Director of the National Institutes of Health, noted to attendees: "I think caBIG™ is a model that I expect to be adapted by other communities, such as those in heart disease and Alzheimer's."

Finally, while the caBIG™ initiative benefited from the participation of numerous patient advocates, caBIG™ itself has not been visible to most cancer patients thus far. It is expected that in the future, patients will benefit from caBIG™ through its ability to facilitate selection of treatment and entry into clinical trials of experimental treatments, monitor for treatment response and adverse effects, and monitor for recurrence of disease.

Dr. John Niederhuber, Director of the National Cancer Institute, predicts that "caBIG™ will drive clinical trials of the future. It will be the way we bring genomics, proteomics, and clinical data together for each patient in a clinical trial."

Chapter 1: Background and Rationale

"We are in the midst of an explosion of knowledge about cancer as a disease process. We are beginning to understand cancer not by what we can see and touch—or by what is revealed under a microscope—but at the molecular level. It is not a question of if, but rather when and how, Molecular Medicine translates into personalized care…

We cannot achieve this (translation) without great interconnectivity and coordination across the cancer enterprise."[16]

The Inception of caBIG™

In 2004, the National Cancer Institute launched the cancer Biomedical Informatics Grid™ (caBIG™) initiative as part of its mission to advance research on cancer and improve clinical outcomes for patients. NCI recognized that the ability to connect people, organizations, and data through information technology would be critical to fulfilling NCI's mission and to taking advantage of the research opportunities offered by 21st century science. caBIG™, overseen by the NCI Center for Bioinformatics (NCICB), began with a three-year Pilot Phase, in order to test the ability of a complex informatics initiative to achieve measurable goals and produce deliverables and to assess the opportunities and challenges of connecting a disparate biomedical community on a national and eventually international scale.

The caBIG™ initiative was an unprecedented effort. At the time of its launch, there was no such interconnected, standardized biomedical informatics platform in place anywhere within the biomedical research community. Even within the academic medical research community, networked data sharing and analysis infrastructure were largely limited to particular departments or internal groups. These disparate groups did not link together all the various research functions in a single institution, much less link entire networks of institutions. In contrast with other national efforts, such as in defense or federally-funded physics research, the nation's biomedical research enterprise had never undertaken such a large Information Technology (IT) project.

[16] *Andrew C. von Eschenbach, M.D., and Kenneth Buetow, Ph.D. (2006) "Cancer Informatics Vision: caBIG™" Cancer Informatics 2006:2 (22-24).*

caBIG™ Pilot Phase Goals

- *Illustrate that a spectrum of Cancer Centers with varying needs and capabilities can be joined in a common grid of shared data, applications, and technologies;*
- *Demonstrate that Cancer Centers, in collaboration with NCI, will develop new enabling software tools and systems to support multiple research organizations;*
- *Demonstrate that Cancer Centers will actively use the grid and realize greater value in their cancer research endeavors by using this network to support powerful collaborations that are dependent on the sharing of data; and*
- *Create an extensive infrastructure that will continue to be expanded and extended to members of the cancer research community beyond the NCI-designated Cancer Centers.*

In that context, the imperatives of the caBIG™ initiative were highly ambitious: to integrate the existing biological and clinical "silos" of cancer research activity; to integrate IT infrastructure, software, and data; and to integrate institutions and people, so that information could be transformed into knowledge at the requisite scale and speed of molecular-based translational research.

Cancer and the Shift to Molecular Medicine

Cancer continues to be a major problem in both the United States and globally, and solutions are still difficult to find. Though progress has been made in the form of declining death rates for certain cancers and increasing 5-year survival rates for many others, 1,500 Americans die every day of cancer, and 1.5 million Americans will hear the words "you have cancer"[17] this year.

An estimated $72 billion[18] is spent on cancer treatment in the United States yearly. As the baby boomers enter their senior years, the number of new cancer cases will increase, thereby increasing the human and economic burden.

At the same time, the pace of basic research discoveries has accelerated. The expansion of scientific knowledge driven by the mapping and sequencing of the human genome, the development of high-throughput technologies for analyzing genes and proteins, and the advancement of systems biology have illuminated cancer as a collection of many individually complex diseases, each with its own molecular signature and characteristics. Such molecular understanding of cancer is being translated into a new generation of individualized diagnostics and therapeutics.

The implications of these new molecular and technological approaches to understanding cancer and other diseases are far-reaching. Twentieth century medicine focused primarily on treatment, attempted to diagnose disease based on morphologic and pathologic analysis of tissues at a cellular level, and did not systematically connect research with clinical care. The era of Molecular ("personalized") Medicine in the 21st century focuses on understanding biological processes

> *"What is required in cancer research to find definitive answers is a system to share data and leverage all the events in the cancer world. It is impossible to succeed without embracing that notion. The concept of caBIG™ is, therefore, right on target."*[19]
>
> Kim Lyerly, M.D.
> Director, Duke Comprehensive Cancer Center

[17] *American Cancer Society, Cancer Fact and Figures 2007.*

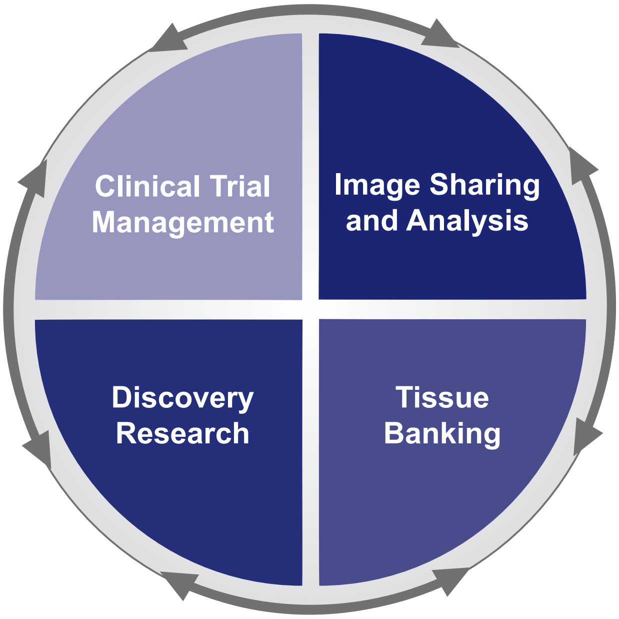
[18] *http://progressreport.cancer.gov/index.asp.*

[19] *Kim Lyerly, M.D., Director, Duke Comprehensive Cancer Center, discussion on May 30, 2007.*

that lead to disease predisposition, initiation, and progression; seeks to diagnose disease early, as well as discover and develop therapeutics, based on molecular characterization and biological understanding; and continuously connects research to clinical care and back to research in a seamless loop of treatment and discovery. This paradigm generates massive amounts of electronic data at every step, necessitating a new, systematic approach to IT connectivity. (See Figure 1)

Similarly, translational research—which transforms scientific discoveries arising from laboratory, clinical, or population studies into clinical applications to reduce cancer incidence, morbidity, and mortality[20] —relies on IT connectivity to amass, analyze, and apply information.

Figure 1. Molecular Medicine: Selected Domains Needed to Enable Translational Research



The Challenges of Connectivity

While the era of Molecular Medicine can potentially increase opportunities for scientific breakthroughs and improved care, its implementation has faced serious structural, cultural, and technological challenges. For example, molecular-based research demands large-scale collaboration among scientists, often from different disciplines and at different institutions. Yet researchers traditionally have worked in isolation in intellectual "competition," each discipline communicating with its own specific terminologies.

Genomics-based technological innovations in this era can rapidly generate extremely large amounts of data, and the ability to collect, analyze, share, and integrate such quantities of biological data in real time is a prerequisite to biological understanding. But information technology within the biomedical enterprise has been slow to develop and is rarely connected between laboratories even within a single institution, much less between different institutions. Frequently, the same types of biomedical data are collected by research groups using their own "home grown" information systems that do not base their data models on any kind of widely shared standard. Additionally, there are often no agreed-upon data models or standards within a single discipline, compounding the inability to share data even among those who collect data using the same analytical platform. The result of many such "disconnects" along the continuum of translational research—from laboratory values, to epidemiological data, to clinical records, to biospecimen records, to imaging data, to molecular data—is that vital scientific discoveries are not made, and the pace of progress against cancer is slowed.

Thus, to address the complexities of cancer and these discontinuities of the research process, a 21st century cancer research enterprise requires *interoperability*; that is, access to integrated tools to collect, analyze, and share data in standardized

[20] *http://www.cancer.gov/trwg/TRWG-definition-and-TR-continuum.*

formats. This interoperability is a means to link together all the scientists, clinicians, patients, and other participants so that they can share such standardized information rapidly.

The Vision and Mission of caBIG™

The vision of caBIG™ is to be the information network enabling all constituencies in the cancer community—researchers, clinicians, patients—to share data and knowledge to accelerate the discovery of new approaches to prevention, diagnostics, and therapeutics, which together will improve patient outcomes.

The mission of the caBIG™ initiative is to provide infrastructure for creating, communicating, and sharing bioinformatics tools, data, and research results, while using shared applications, shared data standards, and shared data models, all operating on a cancer community network (caGrid). Through this infrastructure, caBIG™ supports the development of new types of analysis within and across experiments and allows new forms of collaboration in which biomedical data sets are easily exchanged and more rapidly and efficiently analyzed and integrated via an interoperable set of software tools.[21]

Overarching Objectives of caBIG™

The transformation of the cancer research enterprise into a "worldwide web of information, people and institutions"[22] is by definition a long-term endeavor. Thus, the initial objectives of caBIG™ at the highest level were to:

- Connect scientists and practitioners through a shareable, interoperable infrastructure;
- Develop standard rules, a unified architecture, and a common language to more easily share information; and
- Build or adapt tools for collecting, analyzing, integrating, and disseminating information associated with cancer research and care.

The Principles of caBIG™

To achieve these highest level objectives, four fundamental principles were developed to underlie the activities of caBIG™ and to guide all of its operations:[23]

- **Open Access**: Participation in caBIG™ and the products delivered by caBIG™ are open to all, enabling access to tools, data, and infrastructure by the cancer and greater biomedical research communities.
- **Open Development**: Software development projects are assigned to particular participants, but are informed iteratively with multiple opportunities for review, comment, further modification, and development by the caBIG™ community. The materials that are associated with the planning, testing, validation, and deployment of caBIG™ tools and infrastructure are also open to the entire cancer research community.
- **Open Source**: The software code underlying caBIG™ tools developed with the support of the NCI is available to software developers for use and modification. This software is licensed as open source to promote the reuse of existing code, hence optimizing the full benefit of the research dollars spent. However, the open source license is industry-friendly, allowing commercialization of derivative products and fostering industry interest and innovation, while still adhering to the principle of open source for caBIG™-funded activities (See Case Study: caBIG™ License)
- **Federation**: caBIG™ software and standards enable local organizations, such as Cancer Centers, to share data resources with the larger cancer care and research community and to use resources contributed by others. On the grid, these resources can be aggregated from multiple sites to appear as

[21] *caBIG™ Primer, page 5, https://cabig.nci.nih.gov/overview/cabig-primer, accessed June 1, 2007.*

[22] *Dr. Ken Buetow, NCI Associate Director for Bioinformatics and Information Technologies, "caBIG™: Power of Connection," http://cabig.cancer.gov/resources/video.asp, accessed June 1, 2007.*

[23] *caBIG™ Primer p. 7, https://cabig.nci.nih.gov/overview/cabig-primer/, accessed June 1, 2007.*

an integrated research dataset, while the individual resources remain under the control of the local organizations. This strategy of organizing and providing distributed access to locally-managed tools and data is referred to as "federation" and it represents an alternative to centralized large-scale repositories and systems.

These principles are all aimed at ensuring that the broadest possible community can be productively engaged in cancer informatics, that the solutions are built according to the community's needs, and that the community faces the fewest barriers possible when adopting those solutions.

caBIG™ Philosophy and Culture

Under the supervision of the NCI Center for Bioinformatics, the caBIG™ initiative placed heavy emphasis on collaboration with its many constituencies, not only within the cancer research enterprise but also in the larger external environment of public and private biomedical informatics initiatives. Such collaboration was a primary factor in the inclusive nature of caBIG™ activities, reflected in the program values of "open development" and "open access." In fact, among biomedical data sharing initiatives, caBIG™ has distinguished itself by the focus on its constituent communities and its open approach to data sharing and development.

The concept of a "caBIG™ community" surfaced early in the Pilot Phase, and it drove not only the underlying strategy of the initiative but also much of its organization and culture. A plethora of mechanisms were employed to invite, engage, and sustain relationships between participants who previously had not interacted, from the first step of information-gathering, through ongoing personal and electronic interactions. A wide diversity of participants—including informatics experts, clinicians, bench researchers, patient advocates, and senior executives—were welcomed to caBIG™. The tools, knowledge, and expertise of the caBIG™ initiative have been freely shared among them.

Pilot Phase Goals and Summary of Outcomes

The specific goals for the three-year Pilot Phase of caBIG™ were to:

- Illustrate that a spectrum of cancer centers with varying needs and capabilities can be joined in a common communications framework, transmitting shared data from interoperating software applications and technologies.
    - *Outcome: As of June 2007, there were over 190 organizations participating in the caBIG™ community (See Appendix B for complete listing). This community includes 51 Cancer Centers; federal agencies; and academic, not-for-profit, and industry entities, represented by close to 1,000 individuals.*

> *"caBIG™ is the most significant assembly of informatics minds for cancer ever assembled. And I have certainly met many people through it that I would never have had contact with otherwise. In fact… caBIG™ has transformed this community."*[24]
>
> Michael Becich, M.D., Ph.D.
> Chairman and Professor, Department
> of Biomedical Informatics
> University of Pittsburgh School of Medicine

- Demonstrate that Cancer Centers, in collaboration with NCI, can develop new enabling software tools and systems to support multiple research organizations.
    - *Outcome: More than 300 software components have been delivered during the Pilot Phase, including over 40 end-user applications, and a wide range of infrastructure components, such as data standards and software development toolkits. Over 45 biomedical datasets have been delivered in*

[24] Michael Becich, M.D., Ph.D., Chairman and Professor, Department of Biomedical Informatics, UPMC, discussion on June 13, 2007.

caBIG™ compliant formats, derived from clinical and molecular studies, and they are in use at several cancer research sites. These components were derived from over 5,000 analysis and requirement-gathering task orders issued to the community.

Due to the availability of this software at the close of the Pilot Phase in June 2007, activities were under way to provide installation and support services for a wide range of adopting organizations, including a dedicated rollout of key tools, infrastructure, and interoperability framework to NCI Cancer Centers. The future availability of these tools and datasets over caGrid will enable an increasing number of investigators to share knowledge as it emerges.

- Demonstrate that Cancer Centers will actively use the grid and realize greater value in their cancer research endeavors by using this network to support powerful collaborations that are dependent on the sharing of data.
    - *Outcome: When caGrid (the data transmission network upon which caBIG™ works) was launched, six nodes (i.e., connection points where research organizations log onto the caBIG™ system) and seven software tools/services were available. As of June 2007, caGrid included over 85 services being hosted or accessed by over 80 organizations, and it was extensively being used as a testing platform for the many caBIG™ software applications that will be grid-enabled to support the cancer research community in the Enterprise Phase.*
- Create an extensible infrastructure that will continue to be expanded and extended to members of the cancer research community beyond the NCI-designated Cancer Centers.
    - *Outcome: NCICB has actively collaborated with NCI programs (such as SPOREs,[25] The Cancer Genome Atlas,[26] the Cooperative Groups,[27] and Cancer Genetic Markers of Susceptibility[28]); other NIH Institutes and Centers (such as National Institute of Neurological Disorders and Stroke,[29] National Human Genome Research Institute,[30] National Heart, Lung and Blood Institute[31] and National Center for Research Resources[32]); other NIH-funded programs (such as Biomedical Informatics Research Network,[33] and Clinical and Translational Science Awards[34]); other research grid initiatives and other international informatics initiatives (including a formal agreement with the UK's National Cancer Research Institute[35]). Wherever possible, interoperability with such initiatives has been sought, and caBIG™ tools and infrastructure are consistently made available for either adoption or further development.*

Financial Investment

The caBIG™ Pilot was funded by NCICB at the level of $20 million annually for each of the three years from FY 2004 to 2006.

[25] *http://spores.nci.nih.gov/.*

[26] *http://cancergenome.nih.gov/index.asp.*

[27] *http://www.cancer.gov/cancertopics/factsheet/NCI/clinical-trials-cooperative-group.*

[28] *http://cgems.cancer.gov/.*

[29] *http://www.ninds.nih.gov*

[30] *http://www.genome.gov*

[31] *http://www.nhlbi.nih.gov*

[32] *http://www.ncrr.nih.gov/clinical_research_resources/.*

[33] *http://www.nbirn.net/index_ie6.shtm.*

[34] *http://ctsaweb.org*

[35] *http://www.ncri.org.uk/.*

Chapter 2: Strategic Planning and Initiation

Since the caBIG™ initiative was a novel informatics endeavor in the biomedical research field, NCICB leadership needed to define its strategies and organization de novo, adapting the attributes of other large-scale IT initiatives as appropriate.

Strategic Approaches and Practices

The following strategic approaches were defined and developed in the first year, and they guided caBIG™ activities throughout the Pilot Phase:

- **Establish a community of participants** to serve as advisors about cancer research IT needs, as developers and adopters of the infrastructure and tools, and as disseminators of information about caBIG™ to their home institutions. The initiative should encourage and be able to manage the caBIG™ community growth steadily over time, and it should ensure that the community represented a diverse cross-section of disciplines and sectors.
- **Allocate resources** to ensure that, for each of the identified stakeholder needs, an interoperable software tool would be available to link data from diverse scientific and clinical sources and support their Molecular Medicine research activities. An alternative approach—to focus single-mindedly on a small number of software tools in one or two key areas such as clinical trials or imaging—was thought to be less valuable, since seamless connectivity around the entire translational research process was the ultimate objective. It is important to note that this strategy was not intended to develop every software application from scratch; rather, where software applications already existed, they were to be adopted and adapted to be "interoperable" with caBIG™ in order to save time and resources.
- **Recognize legacy IT systems to facilitate adoption.** The initiative recognized the fact that varying levels of investment in IT infrastructure had already taken place within the cancer research community in recent years, and that there would, as a result, be varying paths to caBIG™ adoption. Emphasis has been on interoperability with caBIG™, rather than on urging institutions to "rip and replace" their existing IT capabilities.
- **Balance caBIG™ project management** between top-down guidance from the NCICB and bottom-up input from the grass-roots of the cancer community. In this way, caBIG™ could incorporate the needs and expertise of the community while adhering to NCI's core mission and sustaining efficient project coordination.
- **Leverage academic institutions** for software development in order to keep the development of tools closely tied to the end-user base; however, fund such development efforts via contracts, as opposed to grants, in order to achieve the rigor of timelines and specific deliverables. This approach

was based upon an expectation of Cancer Center capability in developing professional-grade software tools. It was also intended to ensure close collaboration in the development process among different academic research entities so that requirements and specifications would not be too narrowly specific to any one institution.

- **Leverage existing academic and commercial software**, wherever possible, to avoid unnecessary time and expense redeveloping software. This strategic approach presumed that existing software would be developed with sufficient modularity and programming interfaces to support the addition of standardized connections to the grid.
- **Educate the community** on an ongoing basis about the activities and potential benefits of caBIG™ to address cultural barriers to adoption.

Key Participants

At the outset, NCICB realized that the participants in the caBIG™ Pilot Phase would have to include all sectors of the cancer community.

The key participants included:

- **The National Cancer Institute Center for Bioinformatics (NCICB):** NCICB is NCI's strategic and tactical arm for research information management. Its mission is to provide foundational biomedical informatics infrastructure, tools, and data to serve NCI research initiatives and the cancer research community. NCICB's work enables disparate research data across the "bench to bedside continuum" to be integrated and harmonized. As the guiding organization of caBIG™, NCICB represents NCI in all management and operational decisions and facilitates the activities of the community. Through NCICB, NCI has invested resources for administration and management of caBIG™, alleviating the resource burden on the participating community.
- **NCI-designated Cancer Centers:** The NCI Cancer Centers Program supports major academic and research institutions throughout the United States to sustain broad-based, coordinated, interdisciplinary programs in cancer research. There are now 63 NCI-designated Cancer Centers that continue to work toward creating new and innovative approaches to cancer research.[36] caBIG™ was originally conceived and developed with the input of NCI-designated Cancer Centers, and it was later expanded to the larger biomedical community. Representatives from the Cancer Centers have helped shape and continue to help shape caBIG™ priorities by gathering general needs and specific requirements; developing, refining, and testing standards, programming toolkits, and software applications; providing data sets; and setting policy and guidelines.
- **Patient Advocates:** Patient advocates from NCI's CARRA (**C**onsumer **A**dvocates in **R**esearch and **R**elated **A**ctivities) program and from other organizations have been active members of the caBIG™ community since its inception. CARRA was created to integrate the perspective of people affected by cancer into NCI's programs and activities.[37] Each caBIG™ workspace includes a patient advocate "to help ensure that the caBIG™ end product will ultimately benefit the cancer patient by improving patient care and outcomes in the most effective and timely way possible."[38] Patient advocates also contribute valuable expertise from outside the academic community. According to the patient advocate statement of expectations, "It is the expectation of the caBIG™ Patient Advocates that caBIG™ will have a direct impact on the cancer patient's journey from diagnosis through treatment and beyond, by providing the tools necessary to lead to 1) more rapid translation of basic research to the clinic, 2) centralized clinical trial information that is easily accessible to clinicians, and 3) feedback from the patient to the research community.[39]

[36] *http://www.cancer.gov/cancertopics/factsheet/NCI/cancer-centers, accessed June 1, 2007.*

[37] *http://carra.cancer.gov/about/whatiscarra.*

[38] *caBIG™ 2007 Annual Meeting Newcomer's Guide, page 29.*

[39] *http://cancer.gov, Statement of Expectations, Purpose and Goals from the caBIG™ Patient Advocates, accessed June 1, 2007.*

- **General Contractor:** Booz Allen Hamilton (BAH), a global strategy and consulting firm, was chosen through a competitive process to be the NCI general contractor in the caBIG™ Pilot Phase. As the execution arm of caBIG™, BAH was responsible for day-to-day operational management, including:
  - Coordinating the activities of the participants;
  - Negotiating contracts with participating NCI-designated Cancer Centers and other funded participants;
  - Providing a channel for communications regarding guidance and priorities;
  - Providing measures of participant progress;
  - Fostering accountability; and
  - Providing mechanisms for conflict resolution.
- **Industry Partners and Participants:** Members of industry were welcome to participate in caBIG™ activities from the beginning of the initiative. Mid-point in the Pilot Phase, in September 2005, the Industry Partners Meeting strengthened this connection by formally providing opportunities to commercial organizations to participate in development. Organizations at the meeting included information technology companies and large-scale software integrators, pharmaceutical and biotechnology companies, biomedical research tools vendors, and small, specialized ventures. Industry participation has continued since then, and it has ranged from volunteer involvement in workspace teleconferences and face-to-face meetings to funded development of caBIG™ applications. When the program was launched, all funded participants were academic centers; however, as of the end of the Pilot Phase there were eight directly funded industry participants. This number does not include commercial software developers funded under subcontracts from task orders issued to academic centers.

Additional participants in the caBIG™ Pilot Phase have included, and continue to include, other federal agencies such as the U.S. Food and Drug Administration, academic centers, members of the international cancer community, and international standards associations.

Key Roles

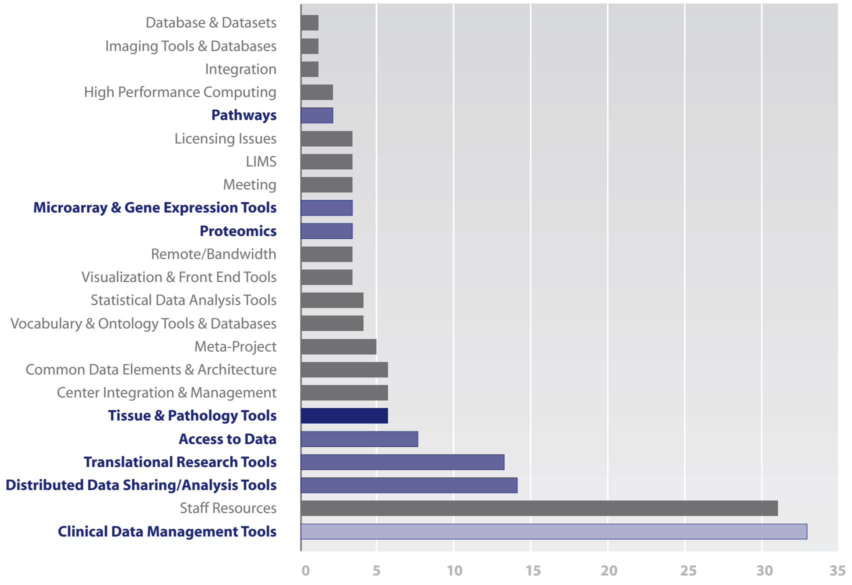
caBIG™ participants played a variety of roles during the Pilot Phase, including:

- **Domain experts:** Workspace participants have contributed to the identification of research community needs, to the prioritization of development of software tools to meet those needs, and to the guidance of the overall direction of the caBIG™ initiative.
- **Developers:** Based upon competitive bidding for contracts, these participants have carried out the work of designing, building, or adapting caBIG™ software tools and infrastructure.
- **Adopters:** These participants—who have a real-world need for a specific application—have acted as beta testers to install and evaluate the software, providing both informatician and basic and clinical researcher feedback on functionality. This information was expected to fuel revisions and improvements to enable use of these tools as the vehicle for providing data to the caBIG™ community.

Assessment of Needs

caBIG™ was to be built by and for the cancer biomedical research community. Each step in the process was stipulated to include open, frequent, and direct dialogue with the community of participants that was identified as central to caBIG™ success. The first step that NCI took in the caBIG™ initiative was to obtain insight from the Cancer Center community to identify critical IT and research needs, as well as their existing strengths and capabilities. NCICB staff and BAH undertook a series of fact-finding trips and other initiatives with Cancer Centers to collect and prioritize informatics gaps. The Cancer Centers expressed a wide variety of needs, with clinical data management, translational research support, specimen management, and data access technology among the most pressing. (See Figure 2) The Centers also expressed varying levels of capability to develop the software applications, tools, and data standards that would fulfill these needs.

Figure 2. Gaining Input from the Cancer Centers



Cancer Center involvement identified priority areas for caBIG™

All of the 49 Cancer Centers that participated in the initial information-gathering activities submitted concepts for pilot projects for review by NCICB. During this review, it was decided that, rather than limiting participation in caBIG™ to 10 to 15 Cancer Centers as initially planned, a strategy of wide inclusion would be adopted, maximizing Cancer Center participation and building on synergies in the respective strengths of the Centers. By the end of October 2003, a list of participating Cancer Centers and their respective roles was finalized, and the caBIG™ Pilot Phase was officially launched at a Kickoff Meeting in Washington, D.C., in February 2004.[40]

Launch Plan and Key Pilot Phase Deliverables

The pilot project concepts submitted by the Cancer Centers were analyzed by NCICB against NCI's own internal needs and assessments. One of the major factors guiding the content of the initial plan was the imperative that the caBIG™ initiative achieve measurable progress during the Pilot Phase. A three-year initial plan was designed that included a set of milestones, defined deliverables, and plans to track program progress against them. (See Figure 3)

The key deliverables fell into several overarching categories, as follows:[41]

- **Software tools:** Software tools are the end-user products that enable investigators to perform their research functions. The list of tools planned for development by caBIG™ was developed in response to the fact-finding needs assessment activities and then refined by the community during the Pilot Phase. caBIG™ software applications have been categorized into the areas of Tissue Banking & Pathology, Integrative Cancer Research, Clinical Trials Management Systems, and *In Vivo* Imaging.
- **Policies and Procedures:** Within the human subjects research and patient protection domains, there is significant regulatory ambiguity that translates to misunderstanding and variability of policies across the biomedical research establishment. Since caBIG™ was expected to develop and deliver tools that manage and

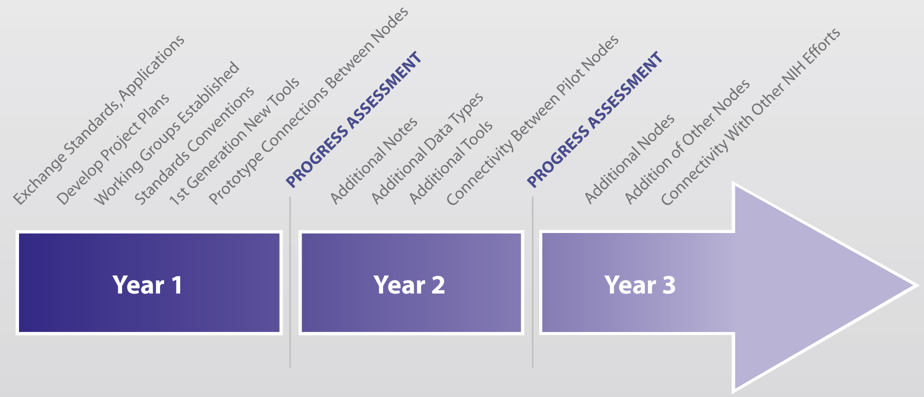
[40] *History of caBIG™ (https://cabig.nci.nih.gov/overview/history), accessed June 2, 2007.*

[41] *caBIG™ community Tools, Infrastructure, Data Resources (https://cabig.nci.nih.gov/inventory/), accessed June 21, 2007.*

transmit patient-derived data, the initiative established the Data Sharing and Intellectual Capital (DSIC) Workspace. This group was tasked with researching these issues and informing the technology developers about requirements that would have to be implemented in software. Additionally, this group has taken on the broader educational task of informing the cancer research community about the ways in which caBIG™ enhances their ability to remain in compliance with regulations and best bioethics practice.

- **The Grid:** caGrid is a set of specifications and software modules that define a data transmission network upon which computer services operate to transmit data between collaborators. caBIG™ software applications were expected to adhere to grid interface specifications, so that they could connect to this network and seamlessly exchange data with other software applications using the same standards. caGrid was designed to include software features that ensure authorization and authentication of users and data security for any service operating on the grid.
- **Standards:** For data to be gathered, stored, and meaningfully exchanged through interoperable software tools, the adoption of standards for data formats, data elements, APIs (Application Programming Interfaces), and vocabularies is required. Such standards ensure that data electronically stored at one institution can be accurately accessed by electronic systems at another institution and incorporated with other relevant data. They also ensure that the language definition ("semantics") used to describe those data can be understood by both machine and human (e.g., that standard names for genes or cancer types are used, and that data are tagged with metadata so that all systems point to the same dictionary defining those terms), so that the information exchanged can be understood and meaningfully integrated.

Figure 3. caBIG™ Initial Plan



- **caBIG™ compatibility evaluation:**[42] This program was developed to define a set of criteria used to measure the extent to which software applications meet caBIG™ compatibility guidelines. The caBIG™ compatibility guidelines define three levels—Bronze, Silver, and Gold—which specify increasingly rigorous concurrence with software functionality, engineering, and documentation standards. As of the end of the Pilot Phase, Bronze and Silver criteria were totally specified, while the Gold level criteria were close to completion. During the Pilot Phase, the cross-cutting Architecture and Vocabularies and Common Data Elements (VCDE) Workspaces implemented a Silver level review process for software developed with funding from the caBIG™ program. In addition, a Bronze compatibility process for any application (whether developed with caBIG™ funding or not) was initiated.

[42] *https://cabig.nci.nih.gov/guidelines_documentation.*

CASE STUDY: caBIG™ LICENSE

Overview of the caBIG™ License

The NCI Center for Bioinformatics gave considerable thought and attention early on in the Pilot Phase to developing policies that would, over time, promote the broadest possible adoption of caBIG™ technology. To that end, NCICB crafted a type of open source software license that takes into account the interests of multiple sectors in the cancer research community, including academics and businesses. The caBIG™ license minimizes all barriers to adoption by essentially eliminating licensing costs and intellectual property restrictions on use of the caBIG™ technology. The caBIG™ license or equivalent must be attached to any software developed using program funds.

Background on Open Source Licenses

Open source licenses for software are predicated on the general availability of source code—the human-readable version of instructions to a computer—so that other developers can understand, use, and modify the software, if desired. However, open source licenses generally have other requirements beyond access to source code. They typically include features such as not restricting what party can distribute software, as well as requiring that the software can be distributed as source code and compiled programs, that the software can be modified or used in derivative works, and that the license must not discriminate against any category of person, entity, or field of use.

Key Benefit of the caBIG™ License

The key benefit of the caBIG™ open source license to developers and distributors of software is that they are free to incorporate caBIG™-developed software into their own products, and they need not release those products' source code. As a result, software developers can more easily develop products that are compatible with caBIG™ (i.e., can interoperate on caGrid) by simply incorporating already developed components that are freely available. Additionally, if they so choose, developers can incorporate caBIG™ technologies, but do so in a way that, while simplifying their own technology development efforts, may result in a product not compatible with caBIG™.

One of the greatest benefits of the non-viral feature of the caBIG™ license accrues to commercial software providers the ability to develop software that is compatible with caBIG™ standards, utilizing freely available open source code, and to subsequently release proprietary products that retain their full intellectual property rights.

Principles Guiding the caBIG™ License Terms

The public health missions of the NIH and the NCI drive the community access requirements for technologies developed with such agencies' funding. NIH policies—coupled with the caBIG™ guiding principles ("open source, open access, open development, and federation")—were the key factors behind development of the license terms to which all NCI-funded caBIG™ developers would have to adhere. These policies and programmatic criteria require that NIH-funded resources developed to support biomedicine, including software and data, should be broadly disseminated to promote research, development, and application; that the broadest possible use of such software and data will ultimately benefit the biomedical community; and that NIH-funded research projects with industry should maintain academic freedom and encourage the broadest possible dissemination of research results.

Key Features of the caBIG™ License

The source code, documentation, and specifications of software developed under the caBIG™ license are available on an open source basis. The principal feature of this license requirement is that the source code and other artifacts are freely available at no charge and without restriction—to any interested party. Additionally, however, the caBIG™ license includes the following key features:

- **The caBIG™ License Is Non-Viral**: Many open source licenses are characterized by a requirement that the open source stipulation propagates to any new software derived from the original software. The industry term for this characteristic is "viral," meaning that the open source stipulation "infects" any new software derived from the original, forcing it also to be open source.

  In contrast, software developed with NCI funding and distributed under the caBIG™ license is "non-viral." Software developers are free to derive new or modified products from caBIG™ software without the requirement to distribute the resulting software products on an open source basis.

- **caBIG™ Software Developed with NCI Funds**: The requirement to release software under the caBIG™ license is specifically dependent upon the source of funds used to develop that software. Basically, if the developer used NCI funds from the caBIG™ program directly or via the caBIG™ general contractor, the developer must release the software under the caBIG™ license terms. If the developer is using other public or private funds, there is no such requirement, thereby providing further flexibility for the software development community.

  NCI developed this policy framework for software access and distribution by negotiating broad rights through the Federal Acquisition Regulation (FAR), which governs procurements by the Federal government. In most cases, if the Federal government contracts for the development of software, the developer retains rights to that software for other uses or sales. In the case of caBIG™, however, as a condition for funding, NCI obtained "unlimited rights" in caBIG™ software developed with such funding, which included the requirement that developers release the source code for caBIG™ products under the caBIG™ license or equivalent terms.

- **caBIG™ Trademark**: It should be noted that simply modifying or incorporating caBIG™ code does not confer the moniker of "caBIG™ compatibility." The NCI has trademarked the term "caBIG™," and it consequently limits the use of the marks to appropriate situations. The caBIG™ license specifically states that end users of caBIG™ licensed code do not obtain the right to use any trademarks owned by NCI in any products except as permitted by NCI or otherwise endorsed by NCI or institutions in the caBIG™ community. NCI will separately license the use of the caBIG™ trademark to those whose software applications have been independently validated as having passed compatibility tests.

Support from the caBIG™ Community

The caBIG™ Data Sharing and Intellectual Capital (DSIC) Workspace includes individuals with legal and licensing backgrounds who were involved in the development of the policies and language of the caBIG™ license. Individuals and institutions that wish to adopt caBIG™ software and need help navigating the intellectual property issues can contact the leader of the DSIC Workspace for general questions about this license structure.

Chapter 3: Programs and Progress

A review of how the caBIG™ initiative unfolded over the first three years reveals stages of activity. The first stage—building community—was critical because it resulted in a community of participants within the "workspace" structure established to manage and inform the project.

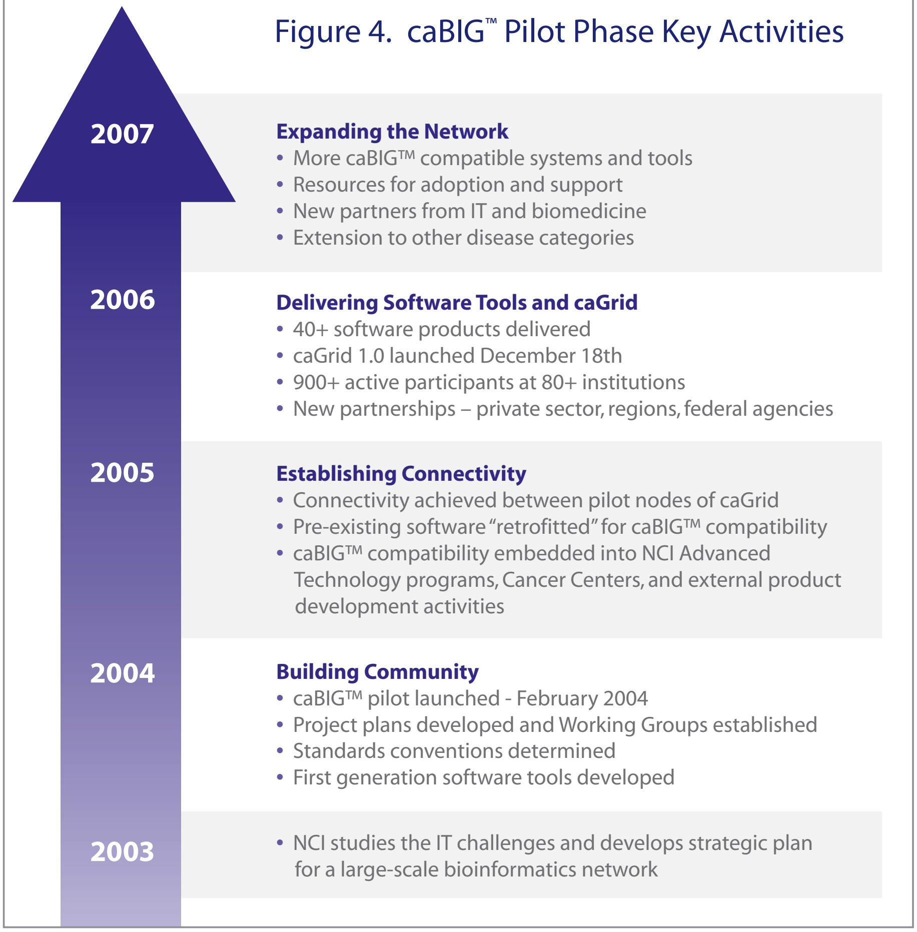


Figure 4. caBIG™ Pilot Phase Key Activities

2007

Expanding the Network

- More caBIG™ compatible systems and tools
- Resources for adoption and support
- New partners from IT and biomedicine
- Extension to other disease categories

2006

Delivering Software Tools and caGrid

- 40+ software products delivered
- caGrid 1.0 launched December 18th
- 900+ active participants at 80+ institutions
- New partnerships – private sector, regions, federal agencies

2005

Establishing Connectivity

- Connectivity achieved between pilot nodes of caGrid
- Pre-existing software "retrofitted" for caBIG™ compatibility
- caBIG™ compatibility embedded into NCI Advanced Technology programs, Cancer Centers, and external product development activities

2004

Building Community

- caBIG™ pilot launched - February 2004
- Project plans developed and Working Groups established
- Standards conventions determined
- First generation software tools developed

2003

- NCI studies the IT challenges and develops strategic plan for a large-scale bioinformatics network

Pilot Phase Stages

As shown in Figure 4, the caBIG™ Pilot Phase progressed through thematic stages of development, in which the number of participants expanded, and their roles evolved over time.

- **Building Community:** In 2003 and 2004, the initiative focused on understanding cancer research needs and on putting the structure and organization in place to enable a multiplicity of development programs. The process of gathering input and determining priorities itself resulted in the formation of a community of individuals and institutions that would carry out the program activities, growing in size and diversity over the course of the Pilot. During this phase, the workspace structure (described below) was created.

- **Establishing Connectivity:** In 2005, the initial software tools and testing sites for the first iteration of caBIG™ tools and infrastructure were established. NCI also established caBIG™ interoperability as the standard for the Institute's leading research and advanced technology initiatives. These initial deployments of caBIG™ technology were key milestones that instructed future software developments, standards, and infrastructure.
- **Delivering Software Tools and caGrid:** During the second year, participants began to focus more on delivery of software applications, data standards, and tools. By the culmination of the Pilot Phase in 2007, over 40 software tools had been developed that span all the research and infrastructure domains identified as areas of focus for caBIG™. caGrid, the essential data exchange network of caBIG™, was launched with six institutions operating on it.

The Pilot Phase was officially completed in March 2007.

> *"The caBIG™ Pilot Phase was extremely ambitious. In three years, it has been trying to accomplish what other endeavors have done over decades."*[43]
>
> Kim Lyerly, M.D.
> Director, Duke Comprehensive Cancer Center

Workspaces

NCICB and BAH organized caBIG™ Pilot Phase activities based on the categories of unmet needs articulated by the Cancer Centers. The Pilot Phase participants were grouped into workspaces, aligned with these categories, that functioned internally as the operational units of caBIG™ and faced outward to the communities that they represented.

Each workspace, which met regularly in person and by teleconference, had a particular area of focus and included teams of Cancer Center representatives provided to the program by the Cancer Center Directors; contractors; both funded and volunteer participants; software developers and adopters; and subject matter (domain) experts who gave direction to the project(s) housed within each workspace. Additionally, several workspaces (especially VCDE and Architecture) were places where mentors could be trained to provide guidance and assistance to the program at large. In each workspace, software technologies that supported similar biomedical research activities were grouped together to facilitate the gathering of information and management of feedback from development efforts.

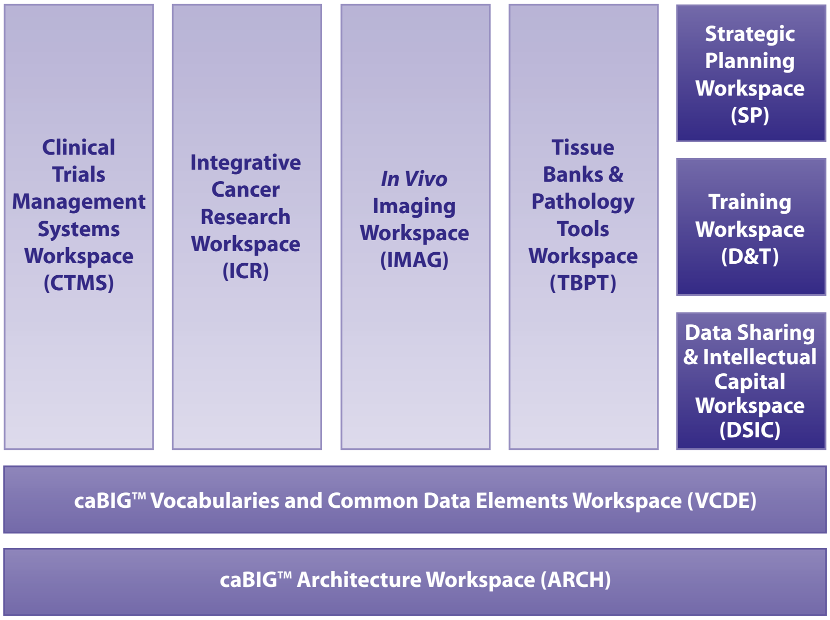
The primary role of each workspace was to determine priorities within that workspace domain and to plan projects based upon caBIG™ strategic priorities. The funded participants in each workspace were also specifically tasked to act as liaisons to ensure a continuous flow of shared information about related activities among all workspaces. Also, special interest groups (SIGs) evolved within workspaces to serve as stewards for particular projects, as well as to keep apprised of events and developments in external standards organizations that could impact workspace activity.

These workspaces were the fundamental building blocks of what grew into the caBIG™ community, and their activity level grew steadily over time. In the first year, the workspaces met for 182 teleconferences; in years 2 and 3 there were almost 500 such teleconferences per year. Workspace attendance has also steadily grown.

[43] *Kim Lyerly, M.D., Director, Duke Comprehensive Cancer Center, discussion on May 30, 2007.*

Domain Workspaces: These technical workspaces were formed in response to unmet software needs identified by 49 NCI-designated Cancer Centers polled during the planning segment of the Pilot Phase, and they were aligned to collect requirements and provide feedback and support for the development and/or modification of a full range of software tools. The workspaces were organized into four categories: clinical trials management systems, integrative cancer research (i.e., integration of molecular and other data from diverse technologies), *in vivo* imaging, and tissue (i.e., biospecimen) banking and pathology tools.

Figure 5. caBIG™ Workspace Organization[44]


Clinical Trials Management Systems Workspace (CTMS)
Integrative Cancer Research Workspace (ICR)
In Vivo Imaging Workspace (IMAG)
Tissue Banks & Pathology Tools Workspace (TBPT)
Strategic Planning Workspace (SP)
Training Workspace (D&T)
Data Sharing & Intellectual Capital Workspace (DSIC)
caBIG™ Vocabularies and Common Data Elements Workspace (VCDE)
caBIG™ Architecture Workspace (ARCH)

As depicted above, the workspaces comprised three broad types of activities: Domains, Cross-Cutting, and Strategic Level.

- **Clinical Trials Management Systems (CTMS):** *CTMS tools are designed to meet the diverse clinical trials management challenges of the Cancer Center community.* Cancer Centers were grappling with an overall lack of clinical trials management systems—and/or diversity of non-interoperable systems—to manage studies, patient registries and enrollment, informed consent, study calendars, and adverse event reporting. Electronic Data Capture (EDC), a common focus area in Clinical Trials Management, represented only one area of need. Ideally, all these systems would communicate using shared data elements for any data in common and would also integrate with any EDC systems for collecting and managing patient clinical data. The CTMS workspace has been developing systems to address this complete range of needs, and it also has been interacting with vendors interested in establishing caBIG™ compatibility for their products.

- **Integrative Cancer Research (ICR):** *ICR tools enable integration between molecular biomedical informatics applications and data.* Translational research requires the integration of traditional clinical study information with the molecular information derived from high-throughput genomic study platforms. Technologies such as DNA sequencing, polymorphism analysis, DNA modification (e.g., methylation), chromosomal changes (e.g., loss of heterozygosity, copy number variation), and gene activity measures, such as RNA and protein expression analysis, are increasingly used in clinical studies. These technologies generate vast quantities of highly structured data that must be integrated with clinical data. Furthermore, the utility of such integrated datasets is then subject to the

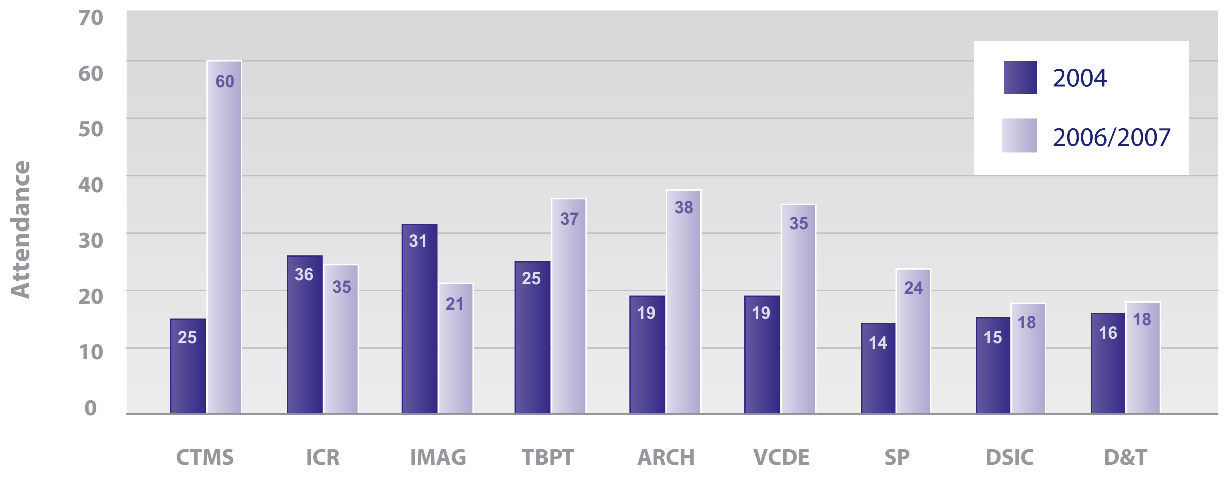
[44] *http://caBIG.nci.nih.gov/index_html/workspaces/index_html, accessed July 3, 2007*

availability of algorithmic and statistical tools that can "crunch" the data in ways useful to the diverse disciplines engaged in translational research. The ICR workspace has been supporting the development of end-user applications and data standards for all the major genomic technology platforms.

- ***In Vivo* Imaging (IMAG):** *IMAG tools and methods manage, analyze, and extract meaning from imaging data, such as X-rays, CT scans, PET scans, and MRIs for both human and animal models of cancer.* Imaging data in biomedical research derives from diverse sources, and it is commonly not organized to support investigations. Among key cancer center needs in utilizing image data for research are standard vocabularies for annotation, image markup tools, applications for extracting de-identified structured data from radiology reports, reference data sets of images, and ways to normalize data obtained from different instruments. The Imaging Workspace, launched in 2005 and comprising both academic users and industry device vendors, had been assessing the Cancer Center needs and prioritizing the most urgent software and standards development effort, initially focusing on annotation, markup tools, and grid connectivity.

- **Tissue Banks & Pathology Tools (TBPT):** *TBPT tools manage the process of collecting, tracking, storing, processing, and distributing tissue samples and their derivatives.* Current genome analysis technologies can generate significant insight when applied to molecular analytes extracted from well-annotated, high-quality tissue samples. Robust systems to collect, manage, and annotate such tissues is a well-known unmet need in cancer clinical research, and future Molecular Medicine is predicated upon meeting that need. Additionally, samples must be annotated with diverse information ranging from the ethical and protocol parameters (e.g., informed consent or clinical trial protocol) under which the donor consented to provide material, to the clinical and diagnostic pathology data collected as part of the donor's clinical care. Such systems must be able to work over the Internet and link researchers managing geographically dispersed

Figure 6. Representative Attendance on caBIG™ Workspace Teleconferences 2004 through 2006/2007



collection protocols and repositories. To support protocols with regulatory requirements, systems must enable a biorepository to track the history of a given specimen and all of its derivatives. TPBT has supported the development of several tools in this space, including biospecimen management, annotation, extraction of structured data from free text pathology records, and de-identification of such data to enable sharing on the grid.

Cross-Cutting Workspaces: These technical workspaces were created to support the specification of the data standards and standard software infrastructure needed by the Domain Workspaces to ensure interoperability of data managed by those respective tools and systems. This requirement for "semantic and syntactic interoperability" is the key technical requirement that needs to be in place for the vision of caBIG™ to succeed. These Cross-Cutting Workspaces have provided a management framework to ensure that such interoperability was built into systems created as part of the caBIG™ program. The Cross-Cutting Workspaces, through multiple channels of communication, ensure that the software tools developed by the Domain Workspaces employ compatible vocabularies, that data formats and application programming interfaces (APIs) are standardized for efficient interchange between different software tools and computer systems, and that the connectivity infrastructure (i.e., caGrid) is in place for data sets to be accessed by researchers at different places, both by geography and by where they work on the translational research process.

- **Architecture (ARCH):** *ARCH* is *software, architecture, and standards for caBIG™ infrastructure, including software development toolkits, application programming interfaces, and the grid network layer*. ARCH guides the development of caGrid, the underlying network architecture and platform that provides the basis for connectivity of software applications and databases, enabling data sharing among caBIG™ participants. The Grid also supports access to data and analytical services, and it provides a system for authenticating users and ensuring permissible and secure access to data made available on the network.
- **Vocabularies and Common Data Elements (VCDE):** *VCDE evaluates and integrates systems and standards for developing, harmonizing, approving, and adopting vocabularies and common data elements, a common standard for defining the biomedical research data managed and transmitted by caBIG™*. A key feature of the way data are structured in caBIG™ is that all specifications include both specific human-readable definitions and machine-readable components that assist software to transmit and incorporate the data without significant programmer customizations. Earlier attempts to create common vocabularies between disciplines focused solely on the human comprehension and communication. Within caBIG™, making the data specification machine-readable provides a framework for rapid adoption of data standards by new software and database systems with a minimum of human intervention.

Strategic-Level Workspaces: These planning and management workspaces develop policies and guidelines that support the other workspaces, and develop and refine the caBIG™ strategic plan.

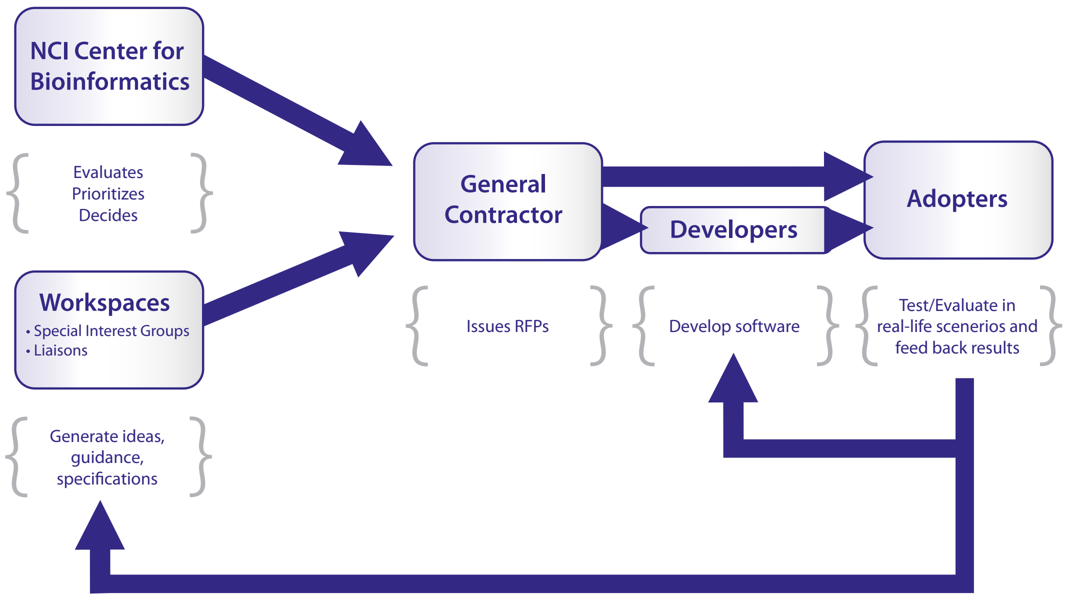
- **Data Sharing & Intellectual Capital (DSIC):** *DSIC addresses issues and develops recommendations related to data sharing, patient privacy, intellectual capital, security, and other policies.* Bringing together researchers, clinicians, technology transfer specialists, attorneys, policy specialists, patient advocates, bioethicists, and bioinformaticians, the DSIC Workspace facilitates education and reduces barriers to caBIG™ adoption by addressing legal, regulatory, ethical, policy, academic, proprietary, security, and contractual barriers to data exchange. DSIC also functions as a resource for members of the caBIG™ community when projects intersect with regulations related to patient privacy and access to patient data.

- **Documentation & Training (DT):** *DT defines guidelines, processes, templates, and tools for developing consistent software documentation and training materials and for fostering mentoring activities throughout caBIG™.* The DT Workspace supports the widespread adoption of the software tools and standards developed in the Domain and Cross-Cutting Workspaces by setting guidelines for supportive documentation to ensure that the tools can be easily employed.
- **Strategic Planning (SP):** *SP provides broad and timely community input to the NCI and general contractor project management leadership.* The caBIG™ initiative is working in domains that rapidly change with new technological and clinical developments. The SP Workspace originally included individuals attuned to this "pulse" and provided ongoing guidance to the high-level caBIG™ directions. caBIG™ project leadership noted SP Workspace input during strategic planning and prioritization of development activities.

The Development Cycle

As shown in Figure 7, the software development process was designed to iterate from identifying needs, to prioritizing activities, to selection of developers, and then to development. caBIG™ also formally funded beta-testing and evaluation within the community through an "adopter" program designed to include real-world situations within Cancer Centers. Testing results and feature requests would feed back via the workspaces, and they would lead to iterative development.

Figure 7. caBIG™ Software Development Flow of Activities


NCI Center for Bioinformatics
Evaluates
Prioritizes
Decides
Workspaces
• Special Interest Groups
• Liaisons
Generate ideas, guidance, specifications
General Contractor
Issues RFPs
Developers
Develop software
Adopters
Test/Evaluate in real-life scenerios and feed back results

Funding Mechanism

The activities of the caBIG™ developers were funded via firm fixed-price (FFP) contracts negotiated between the successful responders to Requests for Proposals (RFP) (initially just the participating Cancer Centers) and BAH. This mechanism was ultimately chosen to facilitate timely and efficient production of specific deliverables and accelerate the overall pace of caBIG™ development. Contracts were issued competitively based on responses from caBIG™ community participants to RFPs issued by BAH for the development of specific tools or standards based on priorities established by the Domain Workspaces and the caBIG™ strategic plan. The contracts included task orders that usually had two to six month timeframes.

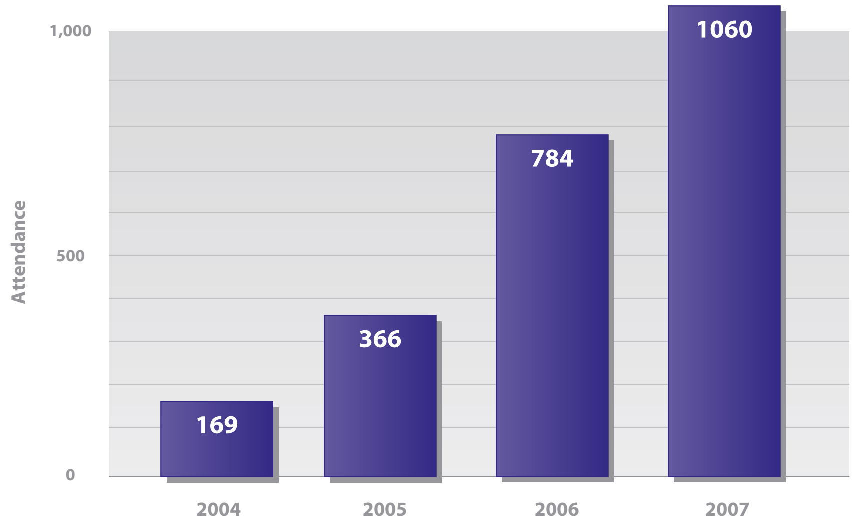
Progress on Deliverables

Virtually all the individuals in the caBIG™ community who shared their observations for this Report believe that the caBIG™ Pilot Phase has exceeded reasonable expectations of what could be done in a three-year period, especially in the context of a rapidly-changing and highly varied biomedical research environment. Many observers pointed out that there were shortcomings and problems, but they also acknowledged that such issues are typical of large IT projects and are precisely what the Pilot Phase was intended to expose and address (See *Hindsight*, Chapter 4).

The progress on deliverables, as well as other accomplishments, included the following key cultural, technical, managerial, and operational successes:

- Formation, launch, and implementation of a vibrant national and international caBIG™ collaborating community, including 54 funded NCI-designated Cancer Centers and voluntary participants from academia, federal agencies, industry, patient advocacy groups, the not-for-profit community, and international cancer institutes. (See Case Study: The caBIG™ Community)
- Establishment of a caBIG™ management structure (under the coordinating supervision of the NCICB) for effective programmatic and operational oversight of the initiative according to industry best practices, governmental policies, and regulations, including a framework to support contracts and funding of caBIG™ project activities. The management structure was designed at the outset to be flexible, with early program staff anticipating that the pilot phase of any large distributed IT project would have to be adaptable to unknown hurdles. Dedicated NCICB and contractor staff over the three years of the Pilot Phase comprised more than 80 professionals.
- Development and deployment of a robust, community-driven organizational and operating structure (the workspaces) designed to capture user needs and priorities identified by stakeholders in the planning phase, refine that

Figure 8. caBIG™ Annual Meeting Attendance 2004 - 2007



information, and ensure that the program stayed in alignment with the vision and goals of the caBIG™ initiative. The caBIG™ Annual Meeting reflected the growth of the community over time: in its first year in 2004, 169 individuals attended; in 2005, 366 attended; in 2006, 784 attended. In 2007, as the Enterprise Phase was launched, over 1,000 individuals attended the meeting. (See Figure 8)

- Launch of development and pilot adoption of individual software products, including component-based clinical trials solutions, tissue bank and pathology tools, and integrative cancer research applications. There are currently approximately 40 tools freely available on the caBIG™ Web site, as well as infrastructure elements and data resources. (See Appendix A for a list and description of caBIG™ tools developed during Pilot Phase.) In 2004 and 2005, a total of 4 tools were released, increasing to 24 tools in 2006. By workspace, this equated to 10 CTMS-related tools, 1 DSIC-related tool, 24 ICR-related tools, 2 Imaging-related tools, and 3 TBPT-related tools developed to the point where they are downloadable via the caBIG™ Web site. As of June 2007, 12 tools have completed the Silver level compatibility review process, and 2 tools have been certified as Bronze compliant.

Figure 9. caBIG™ Deliverables by Category

	2004/05	2006	2007 *(June 30)*	Total
Software Code	293	130	79	**502**
Analysis / Documentation	2830	1452	1041	**5323**
Dataset	9	22	14	**45**
Other	1713	803	275	**2791**
Total	**4845**	**2407**	**1409**	**8661**

- Development of standard data models, relevant common data elements and terminologies for use by the caBIG™ community, as well as processes to expand, adopt, and modify these standards as part of caBIG™ activities.

- Development and promulgation of common guidelines and standards for programming interfaces for caBIG™ application and database network connectivity, including formation of a prototype unifying network architecture for caGrid. Initial demonstrations of connectivity between distinct software applications have been successful. As of June 2007, there were 86 registered services on caGrid, being hosted or accessed by 82 different organizations.

- Development of an understanding among academic developers of the value of professional software engineering best practices for defining, designing, and implementing large-scale solutions involving complex technical and business architectures.

- Coordination and collaboration with related public and private sector healthcare, cancer, and biomedical research IT initiatives.

caBIG™ Connections to Other NCI Programs

The caBIG™ initiative has, since its inception, been connected to and supportive of other NCI programs. Among the NCI Advanced Technology Initiatives for which caBIG™ provides tools and infrastructure are **The Cancer Genome Atlas (TCGA)** (*http://cancergenome.nih.gov*); **The Integrative Cancer Biology Program (ICBP)** (*http://icbp.nci.nih.gov*); **The NCI Alliance for Nanotechnology in Cancer** (*http://nano.cancer.gov*); **Clinical Proteomic Technologies Initiative in Cancer (CPTI)** (*http://proteomics.cancer.gov*); and **Office of Biorepositories and Biospecimen Research (OBBR)** (*http://biospecimens.cancer.gov*).

> *"NCICB and caBIG™ have been terrific partners to FDA, recognizing our role downstream of the research community and working in a very collaborative way."*[45]
> Janet Woodcock, M.D.
> Chief Medical Officer
> U.S. Food and Drug Administration

caBIG™ has also reached out to SPOREs (Specialized Programs of Research Excellence). For example, it now provides a supporting informatics platform for the Prostate and Breast Cancer SPOREs. Other SPOREs, such as the Melanoma and Lymphoma SPORE, are evaluating aspects of caBIG™ software to support various parts of their efforts. For Cooperative Groups (groups of researchers, Cancer Centers, and community doctors who are involved in studies of new cancer treatment, prevention, early detection, quality of life, and rehabilitation), caBIG™ has helped to evaluate which clinical trials management system might be best suited to their needs. The Cooperative Groups have also included a caBIG™ compatibility criteria in their software evaluation process. (*http://cancer.gov*)

caBIG™ Connections to External Programs and Organizations

NCICB has emphasized a collaborative approach toward other federal institutes and agencies as well as toward the international community, with a willingness to share standards, tools, knowledge and experience. Whenever possible, the goal is to achieve interoperability between caBIG™ and other initiatives to facilitate data-sharing across the widest possible biomedical network. Key examples of such interactions include:

- **U.S. Food and Drug Administration:** NCI and FDA have launched several projects that focus on interoperability for regulatory information exchange. The data standardization features of

Program Spotlight: Regulatory Data Exchange (RDE) Initiative

A major opportunity to benefit from establishment of information standards in biomedical research would come in the area of data submission for regulatory review. Under the current system, the majority of submissions to regulatory agencies (such as applications for new drug approvals made to the Food and Drug Administration) are inefficient and almost completely paper-based. This cumbersome process is a contributing factor to the low numbers of new drug approvals despite growth in research spending. A secure and standards-based system for transmitting such data would be a boon to biomedical research, supporting the ultimate goal of speeding research discoveries to patients.

In 2003, the NCI and Food and Drug Administration (FDA) used the Interagency Oncology Task Force effort to launch the Regulatory Data Exchange (RDE) initiative (an outgrowth of Clinical Research Information eXchange, or CRIX) to begin to address some of these problems. The partners to the project have since been expanded to include not only government, but also other key players in the drug discovery, development, testing, and approval process. These key players include industry, academia, standards bodies, and patient advocates. The goal of this group, the caBIG™ Regulatory Data Exchange Steering Committee, is to build a shared, standards-based collaborative research infrastructure for regulatory data and document submission, review and, analysis.

It is intended that the resolution infrastructure will enable the secure transmission of clinical research information among all relevant parties: sponsors, investigators, and regulatory authorities. Additionally, the project should facilitate the adoption of electronic data standards, standardized terminologies, and software systems that perform electronic transactions and submissions. These tools, when made open and accessible to all interested users, are intended to reduce the overall cost of existing information gathering and submissions processes, as well as the tasks of analysis and review.

[45] *Janet Woodcock, M.D., Chief Medical Officer, U.S. Food and Drug Administration, discussion on June 7, 2007.*

FIREBIRD

The identities of the investigators who perform clinical trials represent a key piece of information about any trial data submitted to the FDA for review. These individuals are responsible for all phases of research studies that begin in the laboratory and end with the results of testing in humans being submitted. Their integrity and attestation of results is critical.

The first project launched was the Federal Investigator Registry for Biomedical Informatics Research Data (FIREBIRD) system to manage the registration of investigators, as required by law. The FIREBIRD system permits investigators to register and document their accreditation online with both academic and commercial sector trial sponsors, and it implements a legally enforceable electronic signature capability for documents submitted by the investigator. Among the features of the system is the ability to maintain investigator profiles, manage registration with various entities, submit and receive queries from various entities, and upload documents. The FIREBIRD application provides a "one stop shop" for investigators to submit their information to both trial sponsors and to regulatory agencies, removing the opportunity for ambiguous identity.

As of June 2007, FIREBIRD has been successfully piloted with partners, including NCI, FDA, 5 biopharmaceutical companies, 9 academic *medical centers, over 30 clinical centers and diagnostic labs, and more than 450 investigators. The project is currently in limited production at the NCI Division of Cancer Prevention and is undergoing review to identify enhancements in preparation for adoption by the FDA and NCI community and a broader rollout.*

JANUS

Another key project is Janus, a standards-based clinical data repository that utilizes an open source data model of the same name. Janus was created through a joint venture of the FDA and IBM. The Janus data model provides a standard repository for clinical trial data generated and analyzed during the protocol. Such data range from protocol information (i.e., plans for the trial), clinical data collected during the trial (i.e., actual participant clinical data, outcomes, and adverse events) and analysis plans (i.e., how the data will be reviewed).

During an early pilot phase, a Janus system was populated with artificial clinical trial data to test its ability to meet specific objectives, including support for standards-based data submission, re-usability of analysis tools, reduction of data redundancy, easier data audits, and use of less manual, paper-based data management. In a second phase, the system will be populated with actual (but historical) trial data, to test its ability to support FDA reviewers with their standard suite of data access and statistical analysis tools.

Both pilots have been successful, and Janus is now undergoing a second phase of development that integrates additional industry standard data models.

Role of caBIG™ in the RDE Initiative

There are two key components of caBIG™ that make its tools especially useful for the RDE goal of standardizing data storage and transmission. First, the caBIG™ terminology and data element systems developed by the VCDE workspace support the unambiguous definition of clinical trial information data sets. Second, the caCore and caGrid architectures provide a standards based object-oriented framework, developed with biomedical research in mind, for accessing the data and transmitting it over the Internet.

The long term strategic goal of the partnership is to move these projects out of the government into self-sustaining entities, whose clients would be all the interested parties in clinical trials. As of June 2007, FDA and NCI are soliciting responses from non-governmental organizations to adopt these tools.

caBIG™ are applied to the structured data characteristics desired by the FDA in regulatory submissions.

- **National Institutes of Health:** The NIH Roadmap for Medical Research includes a series of far-reaching initiatives intended to accelerate the pace of life science discovery from laboratory bench to clinical practice, including changes to the clinical research enterprise. Among the Roadmap initiatives are the recently funded Institutional Clinical and Translational Science Awards (CTSAs), a group of programmatic grants specifically designed to support the integration of the various historically independent disciplines in biomedical research that must work together to successfully deliver Molecular Medicine. Another Roadmap initiative in which caBIG™ community researchers have collaborated is the "Re-engineering the Clinical Research Enterprise" information standardization project. caBIG™ has also provided input to the Biomedical Information Science and Technology Initiative (BISTI) as it attempts to recognize the potential benefits to human health that can be realized from applying and advancing the field of biomedical computing.[46]
- **Office of the National Coordinator:** The Office of the National Coordinator for Health Information Technology (ONC) provides counsel to the Secretary of HHS and Departmental leadership for the development and nationwide implementation of an interoperable health information technology infrastructure.[47] NCICB has kept ONC staff updated on caBIG™ activities and resources and collaborated on numerous ONC activities.
- **United Kingdom's National Cancer Research Institute:** NCRI is a partnership between the UK government, charity and industry, which promotes cooperation in cancer research among 20 member organizations. NCI and the NCRI have been cooperating to share caBIG™ tools and achieve data interoperability for sharing of research results. (*http://www.ncri.org.uk/*)
- **Standards Development Organizations:** Organizations such as HL7, LOINC, MGED, and SNOMED develop standard terminologies and data structures for the specific biomedical arena that they represent. caBIG™ participants are representatives to such organizations, and the initiative has co-sponsored joint standards development conferences that included standards organizations and industry. Throughout the Pilot Phase, the caBIG™ initiative has attempted to harmonize with their standards to maximize interoperability.

Recognition of caBIG™

During the Pilot Phase, caBIG™ was identified in a wide range of settings and publications as a contributor to enabling connectivity and data sharing within the cancer community.

The NIH National Center for Research Resources published three reports about caBIG™, as follows:

- ***caBIG™ Overview***, May 2006. This report noted: "Acknowledgements 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 mind set of clinicians, researchers, and funding entities working in the life sciences…caBIG™ is building a cohesive community among the clinical cancer research in which this sea change is occurring. It is as much about bringing people together to embrace a fundamental change in how science is conducted as it is about developing the enabling technology."[48]
- ***caBIG™: Opportunities and Challenges for Use Beyond Cancer***, June 2006. This report noted: "The caBIG™ model appears, generally, to be extensible to other domains, but it will need to be further developed to include more tools and processes…Attention needs to be paid to the human and political aspects of information sharing. Researchers will need to be convinced that sharing

[46] *www.BISTI.nih.gov, as of July 4, 2007.*

[47] *http://www.hhs.gov/healthit/onc/mission/; accessed July 6, 2007.*

[48] *http://www.ncrr.nih.gov/publications/informatics/caBIG.pdf.*

Program Spotlight: BRIDG

As noted elsewhere in this Report, caBIG™ technologies have been used in multiparty public-private partnerships to develop software applications that standardize access to and transmission of validated information about clinical trials. For example, the FIREBIRD application is used to manage data about trial investigators that must flow between the investigator, trial sponsor, and the FDA. In FIREBIRD, a key use of caBIG™ technology has been at the interface—specifically supporting the movement of the investigator data from entity to another. Another area where caBIG™ technologies have delivered value is in standardizing the structure and meaning of the data itself.

When conducting a clinical trial, much attention is focused on data generated about the patients in the trial: their clinical histories, diagnoses, and certainly their responses to the investigative treatment and eventual outcomes. However, one of the more intractable information management domains in clinical trials has been dealing with information about the clinical trial process itself; for example, data about sponsors (the entities that pay for and organize trials), activities to be undertaken during the trial (such as the randomization of patients), or the sites where the trial is being run. The Biomedical Research Integrated Domain Group (BRIDG) Model is a multi-party public-private partnership to build a standard data model that captures this "metadata" about a clinical trial. Written more formally, the BRIDG model is a standard data structure that describes pre-clinical and clinical research, a domain succinctly defined on the caBIG™ Web site as:

> Protocol-driven research and its associated regulatory artifacts, i.e., the data, organization, resources, rules, and processes involved in the formal assessment of the utility, impact, or other pharmacological, physiological, or psychological effects of a drug, procedure, process, or device on a human, animal, or other biologic subject or substance plus all associated regulatory artifacts required for or derived from this effort.[49]

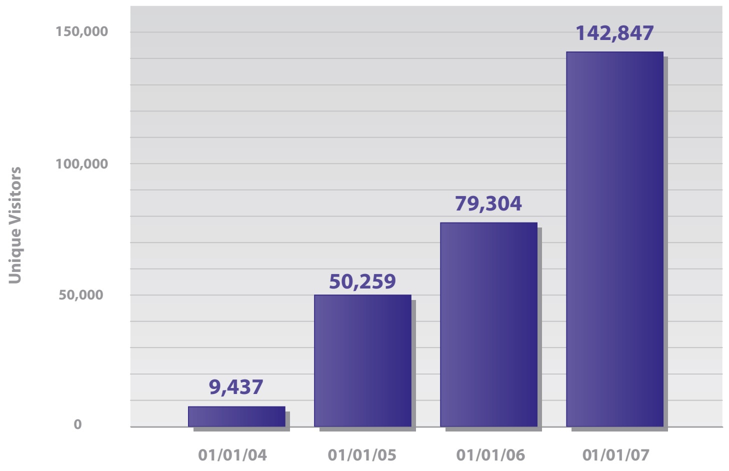
The BRIDG Model emerged in 2003 from a collaborative effort among clinical trial experts from the Clinical Data Interchange Standards Consortium (CDISC, the data standards group for the biopharmaceutical industry), the NCI, the FDA, and Health Level Seven (HL7, the standards development organization for healthcare data).

caBIG™ technologies are extremely well suited to the developing BRIDG Model, because it is intended to be a technology independent and machine readable data structure describing a clinical trial. The caBIG™ policies and tools that specify controlled terminologies, data element structure, data models, and computable metadata about those data elements are all openly developed, made freely available, and provide a pre-made framework for an effort like BRIDG. Accordingly, the caBIG™ program has been a key partner and supporter of BRIDG and was instrumental in the process of bringing the interested parties together.

As of June 2007, version 1.0, the first official release of BRIDG, has been made available to the community and can be freely downloaded from the caBIG™ software repository. Currently, there are several sub-projects in BRIDG, all under active development, including a clinical trial design model, a statistical model (supporting, for example, trial size justification power calculations), and a trial registry model (e.g., for depositing trial existence information in a central registry).

[49] *http://www.bridgmodel.org.*

Figure 10. caBIG™ Web site Visits



data will not compromise the integrity of their studies and that other researchers will not "beat them to the punch" or adversely affect their ability to publish their work. All participants will need to be convinced that data and personal information are secure."[50]

- ***caBIG-Plus™ Conceptual View: Beyond Cancer***, July 2006. This report concluded: "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."[51]

caBIG™ in the Literature

More than 40 peer reviewed papers about caBIG™ or research using caBIG™ tools and resources have appeared in the scientific literature since 2003,[52] as well as abstracts, and news articles in the life sciences press. In addition, numerous presentations about caBIG™ have been delivered at technical, scientific, and research conferences in the United States and internationally.

***Computerworld* Honors Program,** June 2006. In June 2006, the caBIG™ initiative received special recognition by being selected as an Honors Program recipient by *Computerworld*, the leading information technology trade publication. The *Computerworld* Honors Program "annually identifies and records the accomplishments of the men and women, organizations and institutions that are creating the global best practices in leading the world's ongoing IT revolution."

caBIG™ was listed as a "Noteworthy Case Study" in the Journal of the Computerworld Information Technology Awards Foundation, which concluded that: "With a committed team, and by providing mechanisms from the beginning to integrate and drive the program, the caBIG™ program has successfully met the challenges, both technical and social, to create an integrated grid with which cancer research data can be shared broadly throughout the community."[53]

> *"caBIG™ was designed flexibly enough to ask who in the community can contribute, and allow anyone to answer."*[54]
>
> Joel Saltz, M.D., Ph.D.
> Professor and Chair, Department of
> Biomedical Informatics
> The Ohio State University College of Medicine

[50] *http://www.ncrr.nih.gov/publications/informatics/caBIG_OpportunitiesAndChallenges_12-26-06.pdf.*

[51] *http://www.ncrr.nih.gov/publications/informatics/caBIG-Plus_ConceptualView_12-26-06.pdf.*

[52] *https://cabig.nci.nih.gov/Library/Library/caBIG_Scientific_Pubs.html.*

[53] *http://www.cwhonors.org.*

[54] *Joel Saltz, M.D., Ph.D., Professor and Chair, Department of Biomedical Informatics, The Ohio State University College of Medicine, discussion on May 23, 2007.*

CASE STUDY: THE caBIG™ COMMUNITY

Many observers have noted that prior to the inception of the caBIG™ initiative, the cancer community, with just a few prominent exceptions, was almost totally "disconnected"—one researcher from another, one research study from another, one research institution from another. The reasons for that phenomenon were not related simply to information technology—although the investments needed to achieve connectivity of software and data resources can be substantial—but also originated in the history and culture of science and in how individual investigator achievements are rewarded. Virtually all caBIG™ participants and external observers have commented that the formation and cultivation of the caBIG™ community has been its most substantive and potentially highest-impact accomplishment to date.

"The community aspect of caBIG™ has been great."[55]

Robert Beck, M.D.
Vice President and Chief Information Officer
Deputy Director, Population Sciences
Fox Chase Cancer Center

Changing Scientific Strategies

Traditionally, research investigation has been conducted as a solitary endeavor, and projects were funded via grants to individual investigators. Incentives for researchers were in the form of publication in peer-reviewed journals, authorship of which was highly prized and substantially contributed to promotion and ability to attract and direct research funds. In that environment, connectivity among researchers was not perceived as either necessary or desirable.

"Building the community was key to caBIG™success."[56]

Anna D. Barker, Ph.D.
Deputy Director, Strategic Scientific Initiatives
National Cancer Institute

Today, however, molecular-based translational research demands that information be carried seamlessly from one department to another within an institution, or between multiple institutions. Team science has thus become a requisite, as is rapid access to data sets from other researchers' efforts, in a form that can be read, studied, and manipulated. In this paradigm, IT connectivity becomes the lifeblood of the research endeavor. caBIG™ enables this new, more collaborative form of research while at the same time empowering richer forms of traditional institutional research.

The Development of the caBIG™ Community

The concept of gathering supporters arose at the outset of the planning process for the caBIG™ initiative, and continued throughout the Pilot Phase. The overall sensibility was to accept all interested participants; as a result, while the caBIG™ community was largely comprised of informatics experts from NCI-designated Cancer Centers, it also sought, with mixed success, to encompass clinicians, biologists, pathologists, and professionals from other disciplines. The formation of the caBIG™ community was a proactive effort and well resourced: of the 247 funded task orders issued by the general contractor during the first two years, 135 were for support of participation in workspaces. During the third year, that proportion rose to 172 out of 232 task orders.

To cultivate a team spirit and common sense of purpose, the caBIG™ initiative emphasized communications through a caBIG™ community Web site, electronic announcements, face-to-face meetings, teleconferences, annual meetings, newcomer training sessions, conference presentations, and program update documents. As shown in Figure 10, the Web site was a central location for caBIG™ information, and visits have grown steadily over time.

From approximately 100 interested participants at the time of launch in 2004, the caBIG™ community grew to over 1,000 participants from academe, federal agencies, Cancer Centers and related programs, industry, patient advocacy groups, the not-for-profit community, and international cancer institutes. As the initiative entered its Enterprise Phase in 2007, over 190 institutions were participating.

[55] *Robert Beck, M.D., Vice President and Chief Information Officer, Deputy Director, Population Sciences, Fox Chase Cancer Center, discussion on June 1, 2007.*

[56] *Anna D. Barker, Ph.D., Deputy Director, Strategic Scientific Initiatives, National Cancer Institute, discussion on May 16, 2007.*

Chapter 4: Hindsight

As noted in Chapter 1, the caBIG™ initiative was a pioneering endeavor for the biomedical community and, as such, had no directly comparable models on which to base its structure, organization, or operations. (One caBIG™ participant noted ironically that had the caBIG™ concept been presented to an NCI study section as a grant application, it would have been rejected as being overly ambitious.)

With the benefit of more than four years of experience, perspective has now been gained on the cultural, technical, managerial, and operational issues that arose in the Pilot Phase. As described in Chapter 3, NCICB designed its managerial approach to maximize flexibility, so that as it became obvious that some tactics would be more effective than others, it would be feasible to shift gears.

Based on assessments and insights from NCI and NCICB leadership, the general contractor's management team, cancer community participants, leaders of other government agencies, and academic and commercial researchers, a number of criticisms have been noted, and caBIG™ strategies and programs have been adapted to address them.

Program Management and Community Engagement

As noted in Chapter 2, the initial goal for community involvement was for 10 to 15 Cancer Centers to participate. However, interest in caBIG™ was so much greater than expected (49 Cancer Centers opted to participate initially) that the caBIG™ organization had difficulty maintaining effective engagement among its diverse constituents. As a result, the program was overwhelmed, and delays in funding occurred. In this regard, NCICB wished to allow all comers to participate, but it did not "prepare for success" with an operational plan that envisioned a rapid growth scenario. The community became frustrated. Several participants, while understanding the programmatic value of broadening the number of participants, opined that the number of participating Cancer Centers should have remained in the 10 to 15 range, as the larger number resulted in the dilution of funds, strategic ambiguity, and managerial burden.

Over time, NCICB learned to reduce the impact of unpredictable response levels in a variety of ways, including more narrowly focusing RFPs to reflect the strategic plans and priorities of the workspaces, and asking for letters of intent from interested responders prior to submissions of bids to gauge response rates.

Communications

Some participants expressed the view that caBIG™ communications were "terrific" or "excellent." Others, however, noted that while NCICB disseminated information to the workspaces on the "big picture" of the initiative, there was a lack of communication about what different projects were doing specifically to move the initiative forward.

The following areas for improvement in caBIG™ communications have been identified:

- **Setting expectations:** The majority of community participants did not have experience with large-scale information technology projects, particularly in the life sciences/healthcare sectors where IT infrastructure has not been adopted as aggressively as in other sectors. For example, many did not know how long large-scale information technology projects generally take, what they generally cost, and that it is likely that certain components will not succeed the first time. As a result, many community participants expected that at the end of three years the initiative would release a fully functional, commercial-grade set of "shrink-wrapped" software tools ready for immediate end-user deployment. In addition, some participants noted that there was a reluctance within caBIG™ program management to rapidly address projects that were failing to meet development goal; in a couple of extreme cases this perceived reluctance resulted in a community expectation that a particular software application was ready for use when in fact it was not. Some sectors of the community, who had been expecting delivery of software tools by the close of the Pilot Phase, were discouraged as a result of these communications disconnects.

  NCICB communications currently stress what is to be expected in each program area, and timelines for release of software applications have become more explicit.

- **Communications to different user groups:** caBIG™ communications were of relevance primarily to informatics experts, and they were often issue-specific during the software development process. As a result, the biomedical researchers who were not part of the active caBIG™ development community, but who are expected ultimately to be the end users of caBIG™ software tools, were at times left out of the communications loop. In particular, biomedical participants noted that they had difficulty finding out what software tools were in development that would be applicable to their specific areas of research (e.g., clinical trials management, biobanking), the status of those tools, the features of those tools in terms of how functional they would actually be, and their anticipated release timing. One participant noted, "You need to show me what's available and what it will do. Tell me what is coming and when to expect it."

To address this need, NCICB now provides a Web-based inventory of caBIG™ tools, with descriptions of their status, and it plans to enhance these resources further in coming months. In addition, as part of the caBIG™ Enterprise Phase, a portfolio of outreach activities and materials is being developed to engage researchers and facilitate their adoption of caBIG™ tools.

- **The caBIG™ community Web site (http://cabig.nci.nih.gov):** This Web site was the core vehicle for communications during the Pilot Phase, and it continues to be the central repository for archival and current information about the initiative. The Web site suffered from several drawbacks, however. It was not always updated in a timely way (for example, meeting minutes were often months out of date), and it was primarily targeted to the needs of IT people rather than research users. As archival information expanded, it also became unwieldy to use.

  NCICB has initiated numerous improvements to the Web site, including the use of Plone (a content management system), which automates much of the process for maintaining a dynamic Web site. Extensive user input was solicited and usability testing was conducted in multiple phases to better serve the needs of diverse audiences, with a commitment to regular user testing. Additional enhancements to the site are planned for the caBIG™ Enterprise Phase to facilitate adoption by an expanding number of institutions. In addition, NCICB developed a companion Web site for patients and the public to use (http://cabig.cancer.gov) that presents updates about caBIG™ in lay language, in order to encourage involvement of additional constituencies to get involved in the caBIG™ endeavor.

Business Processes

The development of effective business processes was an important task to ensure that product development proceeded at a smooth pace, and milestones were reached on time.

- **Project Management and Technical Governance:** Technology coordination gaps sometimes occurred among planners and developers in the Workspaces and Special Interest Groups. The workspaces were designed to include liaison personnel who would communicate between groups to ensure technical coordination, but this often did not occur early in the Pilot Phase. These liaison gaps were also apparent in a few situations where separate workspaces were duplicating development efforts (in some cases incompatibly) to similar challenges. As a result, some developers now have to retro-fit their applications to achieve harmonization.

  A combination of formal liaisons, mentoring, joint face-to-face meetings, and large scale harmonization activities are helping to address these issues. In particular, the creation of "backbone information models" and other standard models, such as BRIDG, are providing technology while the caBIG™ mentoring program is supplying cross-cutting expertise to all caBIG™ projects.

- **Timing of business process definition and implementation:** Some participants believe that it would have been helpful had the business practices been more concretely defined at the launch of the initiative, although they acknowledge that it would have been difficult to derive such practices without first experiencing the strengths and weaknesses of the caBIG™ community structure.

  Reflecting lessons learned, the business processes of the caBIG™ program going forward will include a comprehensive refinement of the overall financial and business methodologies to allow more (and more flexible) options for funding caBIG™ activities.

- **Strategic planning:** While the caBIG™ organizational structure included a Strategic Planning (SP) Workspace intended to be a robust environment for high level planning, some participants in that workspace felt that the planning process was not sufficiently transparent and that their suggestions and input were not appropriately encouraged or integrated. Observers of the activities of this workspace noted that participants may have focused too narrowly on the issues of their own institutions, thereby overlooking the strategic issues of the larger research community.

As of June 2007, NCICB has restructured the Strategic Planning process and has invited strategic thinkers from the community to help formulate a process for developing the caBIG™ strategic plan for the next three years, including development of a potential new governance structure.

- **Contracting mechanism:** The shortcoming most commonly noted by participants has been the caBIG™ Firm Fixed Price (FFP) contract mechanism. The NCI-designated Cancer Centers, in particular, found this mechanism daunting due to their nearly exclusive familiarity with grant-based federal funding programs. For example, one participant noted that academic institutions operate with virtually no cash reserve; consequently, their staff needs to be predictably funded. This shift from a grants-based model to a firm-fixed priced contract model was a major cultural challenge for the Cancer Centers. In spite of challenges, the NCICB eventually negotiated a base agreement with each participating Cancer Center, via the general contractor. Tasks were executed within that base agreement in order to avoid an unwieldy process of re-negotiating with each Center for each new activity. The process took several iterations to find the best method to channel financial resources to the Centers while setting success criteria that would enable NCICB to re-direct resources based on success or failure. Drawbacks of the initial approach were that dollar amounts were too small and the time frames too limited to achieve the objectives. Cancer Centers expressed concern over the lengths of the contracts, currently set at 12 months, with task orders of two to six months, in contrast with grants-based funding, which typically provides for a multi-year period of support. The relatively short duration of caBIG™ contracts frequently resulted in unfunded periods for staff who were critical to projects; those staff members were often either reassigned, funded through other mechanisms, or occasionally lost to the project.

  NCICB continues to explore more flexible contracting mechanisms.

- **Data sharing agreements:** Initial adoption agreements assumed that institutions adopting caBIG™ software tools would engage in data sharing within the adopter framework for software testing, since such sharing was central to the caBIG™ vision from the start. Data sharing per se was not defined as a "deliverable," however.

  NCICB now ensures that data sharing is an integral part of the agreements with Cancer Centers, in order to achieve the overarching objective of accelerating and enhancing research across the community.

Software Development

- **Academically-based development:** Numerous participants questioned the strategy of providing resources to academic centers for the development of professional-grade software, since capabilities varied widely among Cancer Centers, with some able to deliver appropriate tools on time, but many finding it difficult, if not impossible, to do so. A key component of NCICB's initial strategy was to utilize academic institutions for the development effort, based on the premise that there was significant capability embedded in those sites and that their existing applications, data, and infrastructure could and would be adopted to avoid unnecessary *de novo* development. NCICB's actual experience, however, was that a Cancer Center's scientific and clinical excellence did not necessarily correlate with its ability to generate commercial-grade software, and that metrics to measure such an ability in advance were lacking. Moreover, at some Centers that had advanced their own IT products, it proved impractical to adapt those tools. They were not developed with appropriate architecture or programming interfaces, and they frequently had "hardcoded" local customizations. Furthermore, much software development was carried out in the absence of a formal software development methodology, resulting in undocumented code in multiple styles from developers who had since left their organizations.

NCICB has put mechanisms in place for more rigorous assessment of software development capabilities of respondents to development RFPs. The initiative has also adopted a Unified Process Framework (UPF), a methodology for standardizing management of complex software projects. Additionally, as a result of the increasing stringency of the review process and an increasingly strong pool of participants as the caBIG™ program grows, the current winning proposals often reflect a partnership of academic participants with domain expertise and commercial software developers that provide high-quality and cost-effective development. Also, the developer learning curve has been reduced with the publication of compatibility guidelines enabling new developers to write code from the outset that meets caBIG™ standards.

- **Software tool applicability and usability:** NCICB envisioned that the development of software tools would be executed through a "pairing" of software developers and researchers, in order to address real-life scientific needs. Toward that end, the workspaces were structured to encompass both IT experts and domain expert "end users." But in practice, initial software tool development efforts focused more on the IT and technological aspects of software design and implementation, rather than on the needs of end users.

  As the Pilot Phase advanced, it became apparent that the domain experts were not sufficiently involved, and that as a result, software development primarily became a reflection of the viewpoints of the IT participants and did not necessarily mesh with the day-to-day operations of a molecular biology laboratory or clinical research department. One caBIG™ participant noted that "Overall, I agree with the caBIG™ mantra, but caBIG™ got off the track a bit in the beginning when it went down the 'techy' trail and was not sensitive to what Cancer Center Directors really want, such as patient registries and ways to report adverse events."

  As noted above, NCICB has sought to correct this imbalance. Current teams have shifted to better combinations of IT and research expertise, and user-centered design is incorporated as appropriate.

- **User interface consistency:** caBIG™ software developers designed interfaces unique to their own philosophies, since no guidance on standardizing the user interfaces of the software tools had been developed for the program. As a result, tools produced by different developers had significantly different interfaces.

  As of June 2007, this issue continues to be problematic. NCICB, working with its contractor and with participants, has developed a style guide. Future RFPs will require adherence to this guide as one of the project deliverables.

- **Software testing:** The caBIG™ adopters program was envisioned as a mechanism to test, both from a technical and end user perspective, the quality of software developed within the initiative. Specifically, the end user group was intended to be comprised of actual basic and clinical research users, with real data and real problems, who would devote time to test software in parallel to production settings. In practice, however, the funded adopters were primarily the technically-oriented Cancer Center staff who were involved with caBIG™ from the workspaces, and they infrequently developed testing relationships with such "real" users.

  The caBIG™ Clinical Trials Management Systems Workspace has funded active clinical researchers to participate in Task Forces, whose responsibilities involve user acceptance testing of software in realistic conditions.

- **Controls for software tool readiness and deployment:** Lacking guidance regarding milestones in software tool development and release announcements, developers released tools that had not yet achieved mature status, but which were still in the beta or even alpha phases of development. Some early adopters who installed these software tools stated that, while they expected versions of software that had limited functionality but were useable, they instead encountered numerous software bugs, leading to confusion about the true maturity of some tools. This problem resulted from the experience of academic developers, who usually write software that they themselves install and manage for a local community of users. These developers did not recognize the significant investment in installation testing and support required to move software offsite.

  Some developers also observed, however, that while they wished to carry development of the tools to completion, caBIG™ program management did not fund the projects long enough to deliver that.

  caBIG™ has implemented more rigorous screens of the software development capabilities of institutions bidding for development tasks, as described above. For some organizations, caBIG™ implemented an education regimen that introduced more formal modern software development methodologies. Additionally, the program is incorporating a set of initiatives into the post-Pilot Phase that will provide longer-term and more stable means of funding software throughout the development life cycle.

Role of the Private/For-profit Sector

Some participants felt that early and significant private sector involvement would have been beneficial to the caBIG™ Pilot Phase as a whole, and that it also would have smoothed the transition from the Pilot Phase to the Enterprise Phase. Some also observed that the project should have had a funding mechanism for inclusion of commercial off-the-shelf software (COTS) products. While a few such products became part of the caBIG™ roster of software applications, there are a significant number of commercially-available applications for clinical trials management and molecular bioinformatics with broad functionality that did not make it in. Some participants believe that, because no mutually agreeable framework for modifying commercial software for caBIG™ compatibility existed at the beginning of the Pilot Phase, a significant amount of effort went into "reinventing the wheel" in caBIG™ tool development.

The overarching goal of caBIG™ is to provide and encourage the development of interoperable tools, not siloed point solutions. The NCICB believed that it was important to identify and adopt interoperability standards in the context of real application development and testing, not in the abstract. NCICB did have a process of outreach to the commercial community, and several vendors who attended caBIG™ Pilot Phase meetings did go on to participate in development activities. However, many commercial entities appeared to be reluctant to participate, and they may have believed that caBIG™ as a pilot endeavor was at too early a stage to be a good business opportunity.

Many existing vendor systems known to the community did not include the key technical requirements needed for the caBIG™ data sharing goal. For example, there was a lack of open and documented programming interfaces to enable semantically interoperable communications or connections to caGrid. Additionally, many products did not capture or return data in a manner that leverages terminology and data structure standards, rendering the contents of such systems less usable by scientific peers at other institutions. Therefore, since adequate COTS products were not generally available due to lack of interest from the vendors, the program turned to a number of open source alternatives to prove out the interoperability framework during the Pilot Phase. At the same time, however, new tools were developed that will enable vendors of commercial products to more easily adapt their systems to become caBIG™ compatible. NCI hopes that an increasing *number of COTS vendors will take advantage of these free tools and add caBIG™ compatibility to their products in the future.*

Cultural Shifts

A majority of participants have commented that a major beneficial cultural shift occurred over the three years of the Pilot Phase, in which the concept of connectivity and informatics interoperability began to be embraced across the cancer research community. Some participants noted, however, that at the pragmatic level of software tools adoption, a considerable amount of organizational change management is required, and that more assistance from NCICB is still needed to reinforce adoption of caBIG™. "You need a team of people to visit and say 'here are the tools, and how can we help you?'" said one participating oncologist. Others noted that there is not yet sufficient "mindshare" within their Cancer Center to drive caBIG™ adoption and that such a shift in sensibility and willingness to change will require dedicated internal champions.

Some of these challenges pertain to the role of IT overall, rather than just caBIG™. For example, the scientific community and the information technology community have typically functioned apart from each other, in silos, with little communication or sense of common purpose. Moreover, within many organizations, information technology is often viewed as a technical service rather than a requisite strategic partner. IT experts may not play a role at the leadership level, making it difficult to approach the challenge of connectivity within an organization's own departments, much less between different organizations, in a comprehensive way. The caBIG™ initiative was not fully prepared to address these environmental factors at the outset, which slowed caBIG™ adoption.

NCICB has noted in recent months that the cultural shifts are accelerating; these shifts are most likely due, in some part, to the demands of molecular-based translational research. "Connectivity" also is becoming more widely accepted as an operational imperative. The caBIG™ initiative, in its Enterprise Phase, is also focusing much more heavily on communicating with all key constituencies within each adopting organization so that caBIG™ becomes an integrated part of overall institutional strategies for robust discovery and clinical research of the future. NCICB has launched a program to facilitate Cancer Center adoption of and compatibility with caBIG™. This program includes NCI support of a senior dedicated program coordinator within the Director's office at participating Cancer Centers, who will serve on site to plan, advocate for, and oversee caBIG™ adoption.

Chapter 5: Future Directions

Building from the successful completion of its Pilot Phase, the caBIG™ initiative is implementing expanded programs in a larger universe of networked organizations that link the entire cancer community.

The Launch of the Enterprise Phase

The overarching objective continues to be caBIG™-enabled connectivity of the people, institutions, and data in the cancer community that will lead to answers for the complex questions of cancer biology and thereby improve patient outcomes.

In the spring of 2007, caBIG™ entered its Enterprise Phase, the goals of which include:

- A systematic rollout of available caBIG™ tools and infrastructure to NCI-designated Cancer Centers;
- Formation of an enterprise support network to diversify and broaden access to caBIG™ knowledge and expertise; and
- Engagement with a broader community of government, academic, and private sector entities that in the aggregate constitute a future "ecosystem" of those who will adopt caBIG™, further develop it, and provide compatible software, services, and support.

caBIG™ Adoption Program

NCICB is collaborating with the NCI Cancer Centers program and others to achieve caBIG™ compatibility throughout the cancer research community. The process begins with a self-assessment within each organization wishing to become caBIG™ compatible. This evaluation is followed by the development of a caBIG™ deployment plan. Organizations will then work with NCI to get connected to caBIG™ through the installation of caBIG™ "bundles" or compatibility products, which form the backbone of key software infrastructure and data sharing policies and practices needed to become caBIG™ compatible. The bundles are described below:

- **caBIG™ Clinical Trials Compatibility Framework:** This bundle includes components to support the conduct of human clinical trials and related types of human subject research.
- **caBIG™ Life Science Distribution:** This bundle includes tools and applications that support a variety of basic and translational research capabilities.
- **Data Sharing and Security Framework:** This bundle is based on the caBIG™ Data Sharing Framework. The Framework, when fully built out, will consist of a set of policies, processes, model agreements, model data sharing plans, and other materials that participating Centers agree to help develop and to adopt as appropriate.

Enterprise Support Network

NCI is augmenting its traditional support of the caBIG™ community with the caBIG™ Enterprise

Support Network, comprised of four distinct, but complementary, programs. In addition to ongoing software tool development, adoption, and workspace participation, these new programs will form a technology and domain expertise support network. In the aggregate, the availability and use of these support services will expedite and increase the integration of caBIG™ technology into scientific and clinical research workflows at cancer and academic medical research centers, as well as in pharmaceutical and biotechnology companies. The programs include:

- **Service Providers: Comprehensive Technical and End-User Support.** Service Providers are third-party organizations that will deliver software application and infrastructure technical support to end users and IT professionals on a fee-for-service basis. These organizations will be designated by caBIG™ to help ensure that recipients of their services are getting the most accurate, up-to-date, and effective support of caBIG™ technology.
- **Knowledge Centers: External-facing Domain Experts.** Knowledge Centers will provide domain-specific expertise within the caBIG™ community and serve as points of contact for education, outreach, training, tool enhancements, and deployment needs to the rest of the community. Each Knowledge Center will focus on a niche technology area and serve as an all-purpose consulting and service resource for that particular technology, serving any institution that seeks help.
- **Program Offices: Internal-facing Institutional caBIG™ Expertise.** Program Offices will be caBIG™ teams that will be established within an individual institution and tasked with facilitating and expediting the adoption of caBIG™ technology in that institution. Such offices will be helpful to institutions that wish to deploy and use caBIG™ technologies but lack the initial staffing to ensure smooth and effective rollouts and longer-term best practice usage.
- **Enterprise Adopter Pilot Program: Software Installation/Deployment Initiatives.** The Enterprise Adopter Pilot Program will provide short-term, but comprehensive, support and services to install and successfully introduce specific caBIG™ applications within selected institutions. It will serve to provide the higher levels of support and additional resources that institutions often need when they introduce new and powerful software tools. The program will initially focus on adopting applications from the TBPT Workspace to support biospecimen annotation and management.

> *"caBIG™ will drive clinical trials of the future. It will be the way we bring genomics, proteomics, and clinical data together for each patient in a clinical trial."*[57]
>
> John Niederhuber, M.D.
> Director, National Cancer Institute

caBIG™ and the Patient Community

As mentioned previously, patient advocates were an integral part of the Pilot Phase community, participating in workspaces and other caBIG™ development activities. Most cancer patients, however, have not seen, experienced, or even been aware of the development of caBIG™, since to date it has been a foundational IT endeavor. In the future, however, as caBIG™ tools are integrated into every step of a patient-centric clinical experience, patients will potentially benefit from caBIG™ through its ability to facilitate selection of treatment and entry into clinical trials of experimental treatments, monitoring for treatment response and adverse effects, and monitoring for recurrence of disease. For example, programs such as caMatch enable patients to find clinical trials for which they would be eligible participants. This 21st century paradigm of molecular-

[57] *John Niederhuber, M.D., Director, National Cancer Institute, discussion on May 30, 2007.*

> *"I think caBIG™ is a model that I expect to be adopted by other communities, such as those in heart disease and Alzheimer's. I would love to see those of you who have experienced caBIG™ engage with others and be a catalyst to break down the silos."*[58]
>
> Elias A. Zerhouni, M.D.
> Director, National Institutes of Health

based translational research will thereby be experienced at the individual patient level, with the intended benefits of better prevention, earlier diagnoses, more effective treatment, and improved outcomes. In a biomedical world connected with caBIG™ and "caBIG™-like" technologies, doctors and patients will be able to make treatment decisions using systematic and evidence-based metrics selecting the personalized therapy that gives the patient the best chance for a positive outcome.

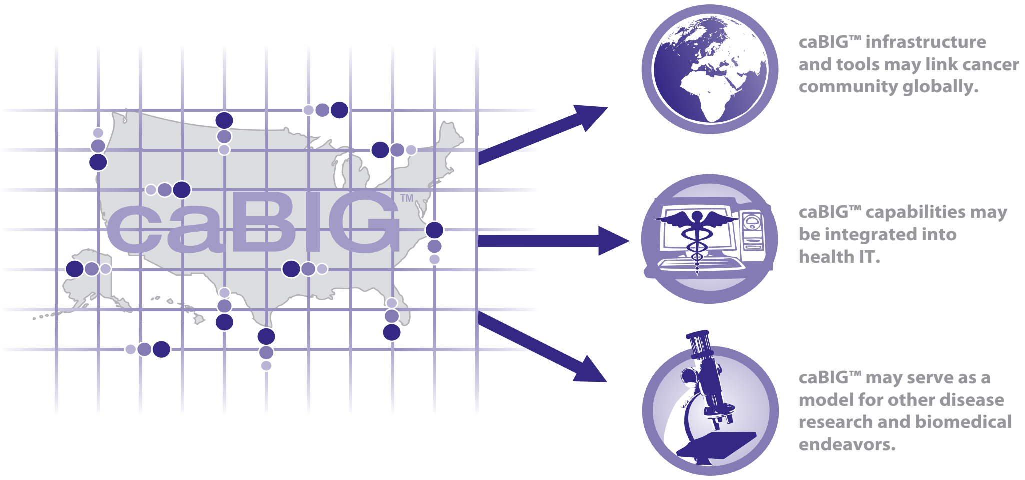
caBIG™ as a Model

Many participants and observers have commented that caBIG™—including its community, tools, vocabularies, standards, and infrastructure—could be expanded beyond the U.S. cancer community to link cancer researchers globally.

In addition, some believe that caBIG™ could serve as a model for other disease areas so that basic and clinical researchers in disciplines such as neurologyand cardiology would not need to start a large-scale IT endeavor *de novo*. Most caBIG™ tools and infrastructure components are widely applicable beyond cancer. Moreover, as the nation's healthcare delivery sector moves increasingly towards electronic health records (EHRs) and systems to connect laboratory, imaging, and clinical data for all patients, it will be logical to use caBIG™ or caBIG™-like technology to link research and clinical care. (See Figure11)

As caBIG™ capabilities spread throughout the cancer community, into other disease areas, and across the translational research continuum, the caBIG™ initiative will have helped to enable a new generation of Molecular Medicine for the benefit of millions of patients.

Figure 11. Future Directions for caBIG™


caBIG™
caBIG™ infrastructure and tools may link cancer community globally.
caBIG™ capabilities may be integrated into health IT.
caBIG™ may serve as a model for other disease research and biomedical endeavors.

[58] *Elias A. Zerhouni, M.D., Director, National Institutes of Health, based on remarks at the 2007 caBIG™ Annual Meeting, February 6, 2007.*

CASE STUDY: EARLY ADOPTION BY DUKE

Duke University, Breast Cancer Clinical Trials, and the Early Adoption of caBIG™

Several departments and centers at Duke University have been actively participating in caBIG™ as a development and adoption site. Within the Comprehensive Cancer Center, the Department of Pathology, and in the inter-departmental Duke Bioinformatics Group, there is significant expertise about the tools developed in most of the caBIG™ workspaces.

Recently, Duke University was selected as the site for a major translational breast cancer study with extensive requirements for tissue banking and molecular profiling. Duke staff's cumulative experience with caBIG™ and the intense informatics support needs of this trial provide a clear opportunity to test the new informatics technologies. Cancer Center leadership has stated that the principles envisioned by the caBIG™ program are requirements for this trial, and that caBIG™ software will be part of the informatics infrastructure. The extent to which caBIG™ software will be part of the informatics infrastructure is the subject of this case study, which describes the Cancer Center's goals for caBIG™, but does so in light of the decisions that early adopters face in selecting any new technology for use in actual research.

> *"We plan to make all the data in this project caBIG™ compatible, and whenever possible, we will make use of caBIG™ software."*[59]
>
> Kim Lyerly, M.D.
> Director, Duke
> Comprehensive
> Cancer Center

The Risks Faced by Early Adopters

With any new technology, early adopters are to be lauded: they face significant risks of increased effort and expense to deploy new technologies, and they also have an obligation to report their experience back to the technology's developers. Such efforts are, however, absolutely critical to finding bugs and, more importantly, to identifying missing functionality that paves the way for that technology to be ultimately successful. In the case of Duke University, Cancer Center leadership completely supports the principles envisioned by the caBIG™ program and has decided to implement all caBIG™ components that are ready now to be integrated in the trial.

Duke's Participation in caBIG™ During the Pilot Phase

Several groups at Duke University have been funded participants in caBIG™ since the initiative's inception. Significant knowledge has been accumulated by Duke groups as both developers and adopters of tools in the architecture, tissue banking, integrative cancer research, and clinical trials workspaces. Duke has specific experience in managing clinical trials data collection using C3D (See Appendix A); using caGrid to link proteomics data generation with an analysis framework, based upon the R statistical language, connected to a graphical user interface; and within the caTRIP project, sharing the data from a large tissue bank management application on the Grid by using caTISSUE as a piece of interface software. (See Appendix A)

[59] *Kym Lyerly, M.D., Director, Duke Comprehensive Cancer Center, discussion on May 30, 2007.*

The Genomics Guided Breast Cancer Trial

The Department of Defense's Breast Cancer Research Program has awarded $6.8 million to Duke to launch a breast cancer research trial in which genomic profiling will be used to guide therapy decisions for women with newly diagnosed cancer. The principal investigator is Paul Marcom, M.D., (Clinical Director, Breast Medical Oncology Program) and the trial will be jointly run in the Comprehensive Cancer Center and the Institute for Genome Sciences and Policy. This trial will rely heavily upon interdisciplinary teams that span basic scientists to clinicians, each of which has its own history of informatics support for its activities.

The trial, since it will test the hypothesis of guiding treatment decisions based upon gene expression profiles from breast cancer tissue, is highly multidisciplinary. In addition to all the clinical trial systems needed for enrollment, electronic data collection, study and patient management, adverse event reporting, and imaging, this trial will also need to make extensive use of annotated tissue bank management systems, as well as pathology informatics and molecular bioinformatics data management and analysis tools. Ideally, the data gathered should be integrated and able to be queried from a number of investigators whose frame of reference could be from different places in the Molecular Medicine translational research process. (See Figure 1)

The caBIG™ platform is being developed to achieve precisely the data integration and interoperability desired for this trial, and there are many caBIG™ capabilities that are extant, especially in data standardization. However, because the trial is likely to begin in the fall of 2007 when many caBIG™ tools do not include the full functionality required by end users, Duke informatics staff will have to make real-time decisions about which caBIG™ tools to use, which to wait for and integrate later, and which to skip in favor of currently available software alternatives.

Feedback to Inform the Future of caBIG™

Since the Duke investigators are launching this breast cancer trial as the caBIG™ initiative is entering its Enterprise Phase, there is an opportunity for significant "real-life" experience. Thus, the Duke efforts to deploy caBIG™ tools for this project may provide the first effective model of supporting a Molecular Medicine translational research program, from end to end, with integrated informatics, and they may provide extensive programmatic feedback to caBIG™ management for future deployment of the platform.

APPENDIX A

caBIG™ Tools and Technologies Developed During the Pilot Phase*

Clinical Software

General Function	Specific Function	Name	Description	URL
Clinical Trials Management	Clinical trial data collection	Cancer Central Clinical Database (C3D)	Cancer Central Clinical Database (C3D) is a clinical trials data management system. C3D collects clinical trial data using standard case report forms (CRFs) based on common data elements (CDEs). C3D utilizes security procedures to protect patient confidentiality and maintain an audit trail as required by FDA regulations. C3D currently supports electronic submission of clinical trials data to the National Cancer Institute's (NCI) Clinical Data System (CDS) and the Clinical Trials Monitoring Service (CTMS/Theradex). C3D consists of three Web-based components: Oracle Clinical, for protocol building; Remote Data Capture, for data entry and management; and Integrated Review / Java Review, for real-time access to clinical data within and across clinical studies to authorized users.	https://cabig.nci.nih.gov/tools/c3d
		Cancer Central Clinical Participant Registry (C3PR)$^{SC(R)}$	Cancer Central Clinical Participant Registry (C3PR) is a Web-based application for managing clinical trial data across multiple cancer clinical trials. The tool is used to improve clinical trials activation and execution by providing a large-scale and efficient Web-based clinical trials information management system available for use by multiple cancer research centers across the country.	https://cabig.nci.nih.gov/tools/c3pr
		Clinical Data Exchange / Lab Integration Hub (caXchange)^{SC}	The Laboratory Integration Hub (caXchange) is an open source software tool used to collect, process, and report laboratory data gathered during a clinical trial.	https://cabig.nci.nih.gov/tools/LabIntegrationHub
		Clinical Trials Object Data System (CTODS)	The Clinical Trials Object Data System (CTODS) is a virtual clinical data warehouse that enables data from any Clinical Trials Data Management System (CDMS) or data source to be available to the cancer research community. It provides clinical researchers with identifiable clinical trials data (as permitted) and provides the broader cancer research community with de-identified clinical trials data (data that have all patient identification information removed).	https://cabig.nci.nih.gov/tools/
	Clinical trial data submission	Clinical Data System (CDS)	The Clinical Data System (CDS) is an independent and stand-alone data submission infrastructure (electronic) that serves as the primary data submission system for NCI-sponsored clinical trials. Data is submitted via a Web-based interface. The system also provides a mechanism for data access by stakeholders, including cancer centers, cooperative groups, and single institutions via a data analysis interface. This interface enables users to view and generate reports about various aspects of the clinical trial process.	https://cabig.nci.nih.gov/tools/CDS
	Patient calendar management	Patient Study Calendar (PSC)	The Patient Study Calendar (PSC) is an open source, standards-compliant software application that can be used by organizations that manage patients on clinical trials. The PSC is a stand-alone, Web-based application providing the ability to create and edit study calendar templates, generate and view prospective calendars of patient activities, track activities as they occur, and manage calendars as they change during a study.	https://cabig.nci.nih.gov/tools/PatientStudyCalendar
	Adverse event management	Cancer Adverse Event Reporting System (caAERS)	The Cancer Adverse Event Reporting System (caAERS) is an open source software tool that is used to collect, process, and report adverse events that occur during clinical trials.	https://cabig.nci.nih.gov/tools/caAERS
	Outcomes analysis	caTRIP$^{SC(R)}$	caTRIP allows users to query across a number of caBIG™ data services, join on common data elements (CDEs), and view their results in a user-friendly interface. Having as its initial focus the enabling of outcomes analysis, caTRIP allows clinicians to query across data from existing patients with similar characteristics to find treatments that were administered with success. In doing so, caTRIP can help inform treatment and improve patient care, as well as enable the search for available tumor tissue, locate patients for clinical trials, and investigate the association between multiple predictors and their corresponding outcomes such as survival.	https://cabig.nci.nih.gov/tools/caTRIP

General Function	Specific Function	Name	Description	URL
Clinical Trials Management	Regulatory and compliance requirement support	Federal Investigator Registry of Biomedical Information Research Data (FIREBIRD)	The Federal Investigator Registry of Biomedical Information Research Data (FIREBIRD) automates the Form 1572 registration process, a key activity in the regulatory data submission process and compliance requirement for investigators participating in clinical trials.	https://cabig.nci.nih.gov/tools/FIREBIRD
Biospecimen Banking	Biospecimen tracking and annotation	caTissue Core^{SC}	caTissue Core is a tissue bank repository tool for biospecimen inventory, tracking, and basic annotation. Version 1.2 of caTissue permits users to track the collection, storage, quality assurance, and distribution of specimens as well as the derivation and aliquotting of new specimens from existing ones (e.g., for DNA analysis). It also allows users to find and request specimens that may then be used in molecular, correlative studies.	https://cabig.nci.nih.gov/tools/catissuecore
	Biospecimen annotation	cancer Text Information Extraction System (caTIES)^{SC}	The cancer Text Information Extraction System (caTIES) is a locator to tissue resources via the extraction of coded information from free text surgical pathology reports. caTIES uses controlled terminologies to populate caBIG™-compliant data structures. It provides researchers with the ability to query, browse, and acquire annotated tissue data and physical material across a network.	https://cabig.nci.nih.gov/tools/caties
		caTissue Clinical Annotation Engine (CAE)	caTissue Clinical Annotation Engine (CAE) is a Web-based user interface for standards-based manual annotation of biospecimens with clinical information. It supports importing structured data from clinical information systems such as anatomic pathology laboratory systems (APLIS), cancer tumor registries, and clinical pathology laboratory systems. caTissue CAE allows the integration of annotations from multiple sources within the cancer centers, providing a complete picture of a patient's disease. The potential users of caTissue CAE include researchers, tissue bankers, pathology assistants, pathologists, and registrars. The current version contains only anatomic pathology annotation.	https://cabig.nci.nih.gov/tools/cae
Image Analysis	Cancer image archive and retrieval	National Cancer Imaging Archive (NCIA)	The National Cancer Imaging Archive (NCIA) is a searchable, national repository integrating *in vivo* cancer images with clinical and genomic data. NCIA provides the cancer research community, industry, and academia with public access to: DICOM images, image markup, annotations, and rich meta data.	https://cabig.nci.nih.gov/tools/NCIA
Data Mapping	Data mapping transformation and validation	caAdapter	caAdapter is an open source tool set that facilitates data mapping, transformation, and validation among different kinds of data sources, including HL7 version 3 messages, Study Data Tabulation Model (SDTM) data sets, object models and data models. It possesses the capability to perform vocabulary validation and integrates with NCICB caCORE components. caAdapter has a component-based architecture that offers a tool set to support data mapping, transformation, and standard data reporting.	https://cabig.nci.nih.gov/tools/caAdapter
Pre-Clinical Data Management	Pre-Clinical data management	Electronic Laboratory Management Information Resource (caELMIR)	The Electronic Laboratory Management Information Resource (caELMIR) provides the pre-clinical scientist with a data management system to record experimental data. caELMIR is a "LIMS" system for basic scientific data.	https://cabig.nci.nih.gov/tools/caelmir

** For explanation of symbols "SC(R)" and "SC" see page 53*

Data Analysis Software

General Function	Specific Function	Name	Description	URL
Data Integration	Molecular biology data analysis	geWorkbench	geWorkbench provides an innovative, open-source software platform for genomic data integration, bringing together analysis and visualization tools for gene expression, sequences, pathways, and other biomedical data. It gives scientists transparent access to a number of external data sources and algorithmic services, combining these with many built-in tools for analysis and visualization (at present more than 40 distinct analysis and visualization modules are part of the platform).	https://cabig.nci.nih.gov/tools/geWorkbench/
		GenePattern	GenePattern puts sophisticated computational methods into the hands of the biomedical research community. A simple application interface gives a broad audience access to a growing repository of analytic tools for genomic data while an Application Programming Interface (API) supports computational biologists. GenePattern is a powerful analysis workflow tool developed to support multidisciplinary genomic research programs and designed to encourage rapid integration of new techniques.	https://cabig.nci.nih.gov/tools/GenePattern
	Molecular biology and clinical data analysis	caIntegrator^{SC}	caIntegrator is a novel translational informatics platform that allows researchers and bioinformaticians to access and analyze clinical and experimental data across multiple clinical trials and studies. The caIntegrator framework provides a mechanism for integrating and aggregating biomedical research data and provides access to a variety of data types (e.g., Immunohistochemistry (IHC), microarray-based gene expression, SNPs, clinical trials data etc.) in a cohesive fashion.	https://cabig.nci.nih.gov/tools/caIntegrator
		caBench-to-Bedside (caB2B)	caBench-to-Bedside (caB2B) is a caGrid client that permits bench scientists, translational researchers, and clinicians to leverage caBIG™ compatible data and analytical services through a graphical user interface. Its metadata-based query interface enables end users to search virtually any caGrid data service. This single tool was designed to integrate and analyze diverse biomedical data sets seamlessly. It has been developed to facilitate the individual steps of cancer research analyses and to reduce the bench-to-bedside barrier.	https://cabig.nci.nih.gov/tools/caB2B
Genome Analysis	Genome annotation	SEED^{SC}	SEED is a tool for making and sharing genomic annotations, and it can be used to access annotations already made, perform queries upon them, and perform computations upon the annotations or upon information collated with the annotations, (e.g., perform a psi-blast amongst sequence data for all genes matching an annotation search). In these two broad categories, the former requires read/write services while the latter requires just read-only services.	https://cabig.nci.nih.gov/tools/SEED
	Genome annotation to find disease-causing mutations	Transcript Annotation Prioritization and Screening System (TrAPSS)^{SC}	Transcript Annotation Prioritization and Screening System (TrAPSS) predicts the potential of gene sub-sequences to contain disease-causing mutations, utilizes annotation to prioritize focused regions of a gene during mutation screening, and aids scientists who are searching for the genetic mutation or mutations that are linked to expression of a disease phenotype.	https://cabig.nci.nih.gov/tools/TrAPSS
	Microarray probe annotation	Function Express (caFE)^{SC}	Analysis of microarray data using gene annotation is essential for the identification of aberrant pathways in tumors. Function Express is an ETL tool that annotates probes on microarrays using publicly available biomedical databases and automatically update these annotations on a regular basis. This data can be queried using a Web-based query interface or programmatically using the caCore-like Application Programming Interface (API).	https://cabig.nci.nih.gov/tools/Function_Express

General Function	Specific Function	Name	Description	URL
Genome Analysis	Determination of biological functions using gene ontology	GoMiner™	GoMiner™ is a tool for biological interpretation of "omic" data—including data from gene expression microarrays. Omic experiments often generate lists of dozens or hundreds of genes that differ in expression between samples, raising the question "What does it all mean biologically?" To answer this question, GoMiner™ leverages the Gene Ontology (GO) to identify the biological processes, functions, and components represented in these lists. Instead of analyzing microarray results with a gene-by-gene approach, GoMiner™ classifies the genes into biologically coherent categories and assesses these categories.	https://cabig.nci.nih.gov/tools/GOMiner
	Mapping and interlinking of genomic dividers	GeneConnect	GeneConnect is a caBIG™ mapping service that makes interoperability possible by interlinking approved genomic identifiers. These include: Ensembl Gene ID, Ensembl Transcript ID, Ensembl Protein ID, Entrez Gene ID, UniGene ID, GenBank mRNA Accession Number, GenBank Protein Accession Number, RefSeq mRNA Accession Number, RefSeq Protein Accession Number, and UniProtKB Primary Accession Number. To interlink all of these identifiers, database annotations (either direct or inferred) and an alignment engine have been used to construct pair wise connections, and then all-to-all relationships have been calculated by traversing all possible combinations of edges in the graph using every node as the starting point. For each query, composed of one or more input identifiers and a set of paths that may be traversed, the Path Score and Frequency are calculated. The GeneConnect application is an independent component with the following modules: GeneConnect server, Web application, and XML-RPC server.	https://cabig.nci.nih.gov/tools/GeneConnect
Statistical Analysis	Multivariate cluster-modeling	Visual Statistical Data Analyzer (VISDA)	The Visual Statistical Data Analyzer (VISDA) is an analytical tool for cluster modeling, visualization, and discovery. Being statistically-principled and visually-insightful, VISDA exploits the human gift for pattern recognition and allows users to discover hidden clustered data structure within high dimensional and complex biomedical data sets. The unique features of VISDA include its hybrid algorithm, robust performance, and "tree of phenotype." With global and local biomarker identification and prediction functionalities, VISDA allows users across the cancer research community to analyze their genomic/proteomic data to define new cancer subtypes based on the gene expression patterns, construct hierarchical trees of multiclass cancer phenotypic composites, or to discover the correlation between cancer statistics and risk factors.	https://cabig.nci.nih.gov/tools/VISDA
	Statistical corrections	Distance Weighted Discrimination (DWD)	Distance Weighted Discrimination (DWD) is a tool that performs statistical corrections to reduce systematic biases resulting from different sources of RNA, batches of microarrays, and particularly different microarray platform.	https://cabig.nci.nih.gov/tools/DWD
Data Services	Protein data collection and analysis	Protein Information Resource (PIR)[SC]	The Protein Information Resource (gridPIR) service provides a data resource for genomic and proteomic research containing rich, high quality, and annotated information on all protein sequences. The resource is supported by UniProt Knowledgebase (UniProtKB) and other relevant Protein Information Resource (PIR) databases.	https://cabig.nci.nih.gov/tools/PIR
	Microarray data collection and analysis	caArray	caArray is an open source microarray data management system that allows users to submit, annotate, and download microarray data. caArray was developed using the caBIG™ compatibility guidelines, as well as the Microarray Gene Expression Data (MGED) society standards for microarray data. Compatibility with these standards and guidelines will facilitate data sharing and integration of diverse data types including clinical, imaging, tissue, and functional genomics data. A number of analytical tools that connect to caArray are already available, including geWorkbench and GenePattern, which both provide a variety of data analysis, visualization, and annotation functions for microarray and other data types.	https://cabig.nci.nih.gov/tools/caArray

Data Analysis Software, Cont'd

General Function	Specific Function	Name	Description	URL
Data Services	Collection of data on animal models of human cancer	Cancer Models Database (caMOD)$^{SC(R)}$	The cancer models database (caMOD) provides information about animal models for human cancer to the public research community. caMOD provides the following key capabilities to its users: Data Submission—Data in caMOD are extracted from the public scientific literature by curators or they are directly submitted by scientists. Search—Users can retrieve information about the making of models, their genetic descriptions, histopathology, images, microarray data, and therapeutic trials in which the models were used, among other things. System Function Administration—The Admin function provides services for user registration, review of submitted models, and database management.	https://cabig.nci.nih.gov/tools/caMOD
Protein Analysis	Mass spectrometry data analysis	RProteomics^{SC}	RProteomics is a package for analyzing mass spectrometry proteomics data. Specifically, these routines have been tailored for the analysis of MALDI style data, including LC-MS data, although use of these routines for the analysis of SELDI data and FT-ICR data is possible. The current implementation of the RProteomics system meets the Gold level of caBIG™ compatibility. This guide is geared toward the use of the application via the GUI. All components should be installed according to the instructions found in the RProteomics System Installation/Administration Reference Manual. Data and analyses may be performed locally or via the grid.	https://cabig.nci.nih.gov/tools/RProteomics
		Q5	Closely associated with RProteomics, Q5 has shown utility in disease classification of expression-dependent proteomic data from mass spectrometry of human serum. Q5's ability to classify complex fragment mixtures was evaluated by testing its ability to discriminate the mass spectra of healthy vs. diseased human serum samples. The two disease states examined in testing were ovarian and prostate cancer. Existing screening methods for both cancers carry a low positive predictive value (PPV).	https://cabig.nci.nih.gov/tools/Q5
	Management of 2D gel lab processing	Proteomics Laboratory Information Management System (protLIMS)^{SC}	protLIMS is a Laboratory Information Management System dedicated to studies in the realm of proteomics. The goal of the prototype version is to develop the system to the point in the analytical workflow where samples are prepared for mass spectroscopy. This entails recording of biological sample data, sample preparation, protein separation/resolution, and isolated protein sample preparation, etc.	https://cabig.nci.nih.gov/tools/Proteomics_LIMS
	Collection and analysis of protein data relevant to cancer	Cancer Molecular Pages (CMP)	The Cancer Molecular Pages (CMP) project was developed to automatically annotate cancer-related proteins in and make the annotations widely available on the caBIG™ grid. CMP has these essential features: appending local annotations to a computed database, processing lists of proteins, using a range of homology tools, and linking protein entries to relevant caBIG™ datasets.	https://cabig.nci.nih.gov/tools/Cancer_Molecular_Pages
		Computational Proteomics Analysis System (CPAS) / LabKey	Computational Proteomics Analysis System (CPAS) is a proteomics application that runs on LabKey Server, the open source platform for high-throughput biology research. CPAS automates the process of peptide scoring and running tools from the Trans Proteomic Pipeline, and then loads the results into a SQL database. Researchers can use the Web interface of CPAS to analyze and compare results across thousands of experiments and share their results securely with colleagues around the world. CPAS is now caBIG™ Silver-level compliant.	https://cabig.nci.nih.gov/tools/cpas
Gene Expression and Protein Analysis	Microarray, DNA processing, and assessment	Biconductor	Bioconductor is an established open source collection of software packages for high-throughput genome analysis. Packages adapted for caBIG™ allow preprocessing of microarray data, DNA copy number assessment from gene expression data, and SELDI-TOFF mass spectrometry peak finding.	https://cabig.nci.nih.gov/tools/Bioconductor

General Function	Specific Function	Name	Description	URL
Pathway Analysis	Analysis of microarray and pathway data relevant to cancer	Quantitative Pathway Analysis in Cancer (QPACA)	Quantitative Pathway Analysis in Cancer (QPACA) is a pathway-based tool that provides a set of routines for quantitative analysis of microarray data in the context of pathways and a set of tools for visualization of pathway structure.	https://cabig.nci.nih.gov/tools/QPACA
	Human pathway analysis	Reactome^{SC}	The Reactome data sharing establishes Reactome as a data feed to caBIG™ by extending the caBIO data model, providing Web Services APIs, and setting up a Web Services server.	https://cabig.nci.nih.gov/tools/Reactome
	Analysis of biomolecular interactions and key cellular processes	Pathway Interaction Database	The Pathway Interaction Database is a highly structured, curated collection of information about known biomolecular interactions and key cellular processes assembled into signaling pathways. Users can query the database by pathway name, by molecule name, or by accession. Molecular detail includes protein post-translational modifications and cellular location. Annotations on interactions include literature citations and evidence codes. All data is available in BioPAX Level 2 export.	https://cabig.nci.nih.gov/inventory/data-resources/pathway-interaction-database/
		Pathway Tools	Pathways is a suite of tools, which include: cPath, an open source pathway database and software suite designed for systems biology research and Cytoscape, used to overlay gene expression data and visualize the results. The third piece is Biological Pathway Exchange language, which is a data standard currently under development to model biological pathways.	https://cabig.nci.nih.gov/tools/Pathways_Tools

Infrastructure

General Function	Specific Function	Name	Description	URL
Core Infrastructure	Data sharing network	caGrid	caGrid is the underlying network architecture that provides the basis for connectivity between all of the cancer community institutions, allowing research groups to tap into the rich collection of emerging cancer research data while supporting their individual investigations. caGrid manages and securely shares information and analytic resources using locally managed access control policies and by using strongly typed data objects in XML format.	https://cabig.nci.nih.gov/inventory/Infrastructure/
	Standardization of clinical data exchange	Biomedical Research Integrated Domain Group (BRIDG) Model	The Biomedical Research Integrated Domain Group (BRIDG) project is a collaborative effort of stakeholders from the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), the National Cancer Institute (NCI), and the U.S. Food and Drug Administration (FDA) to produce a shared view of the dynamic and static semantics that collectively define a shared domain-of-interest, (i.e., the domain of clinical and pre-clinical protocol-driven research and its associated regulatory artifacts).	https://cabig.nci.nih.gov/inventory/infrastructure/bridg/
	Common data management and application development framework	Cancer Common Ontologic Representation Environment (caCORE)	The Cancer Common Ontologic Representation Environment (caCORE) is the open source group of software products developed by the NCI Center for Bioinformatics (NCICB). By providing a common data management and application development framework, caCORE helps streamline informatics development throughout the cancer community. Components of caCORE support the development and utilization of semantically interoperable data systems, ensuring that biomedical research data deployed within this framework is consistent and comparable.	https://cabig.nci.nih.gov/inventory/Infrastructure/
		Cancer Bioinformatic Infrastructure Objects (caBIO)SC	The Cancer Bioinformatic Infrastructure Objects (caBIO) project provides a robust platform and independent infrastructure that illustrates data integration techniques, allowing researchers to perform innovative analysis via a variety of APIs, Web services, and html interfaces. caBIO employs industry-standard software engineering methodologies to develop objects, data models, middleware, vocabularies, and ontologies for biomedical research. caBIO is the primary programming interface to caCORE, a synthesis of software, vocabulary, and metadata models for cancer research. caBIO objects are implemented using Java, and they represent biological and laboratory entities such as genes, chromosomes, sequences, SNPs, libraries, clones, pathways, and ontologies.	https://cabig.nci.nih.gov/tools/cabio
	Development of controlled vocabularies	NCI Enterprise Vocabulary Services (EVS)	The NCI Enterprise Vocabulary Services (EVS) develops standard, controlled vocabularies as part of caCORE. This service produces the NCI Thesaurus and the NCI Metathesaurus, which is based on NLM's Unified Medical Language System Metathesaurus and supplemented with additional cancer-centric vocabulary.	https://cabig.nci.nih.gov/inventory/Infrastructure/
	Standardization of metadata in the form of common data descriptors	Cancer Data Standards Repository (caDSR)	The Cancer Data Standards Repository (caDSR) is a metadata registry in caCORE that stores and manages Common Data Elements (CDEs) that are developed by caBIG™ participants and various NCI-sponsored organizations.	https://cabig.nci.nih.gov/inventory/Infrastructure/
Vocabularies	Creation of a "caCORE-like" software system	caCORE Software Development Kit (caCORE SDK)	The caCORE Software Development Kit (caCORE SDK) is a set of tools designed to aid in the design and creation of a "caCORE-like" software system. This system is "semantically integrated" — all exposed API elements have runtime accessible metadata that defines the meaning of the elements using controlled terminology.	https://cabig.nci.nih.gov/inventory/Infrastructure/

General Function	Specific Function	Name	Description	URL
Vocabularies	Vocabulary installation and publication	LexBIG	The LexBIG vocabulary service represents a compressive set of software and services to load, publish, and access vocabulary. Cancer Centers can use the LexBIG package to install NCI Thesaurus and NCI Metathesaurus content queryable via a rich application programming interface (API). LexBIG services can be used in numerous applications wherever vocabulary content is needed.	https://cabig.nci.nih .gov/inventory/ Infrastructure/
	Standard vocabulary development for human and mouse anatomy	Mouse-Human Anatomy Mapping Ontology (MHAP)	The Mouse-Human Anatomy Project (MHAP) provides a mapping and harmonization of Human and Mouse anatomical descriptors as they are currently used for murine and human models by Mouse Genome Informatics and the NCI Thesaurus. This ontology will facilitate closer integration of human and mouse cancer data, promote the use of the mouse as a model for cancer research, and accelerate translation of basic research discoveries into new clinical therapies.	https://cabig.nci.nih .gov/inventory/ Infrastructure/
	Standard vocabulary development for cancer nutrition	Cancer Nutrition Ontology	The Cancer Nutrition Ontology provides a publicly available, unified set of nutrition vocabularies, based on external standards, that would provide vocabulary/ conceptual uniformity across applications. The need for a nutrition ontology is driven from studies that search for nutritional factors that alter the risk of getting cancer. Clinical trials study chemopreventative agents and primary or adjuvant therapeutic agents, such as SELECT (Selenium, Vitamin E, and prostate cancer). Development of this ontology involved gathering input from numerous experts and sources, including USDA, InFoods, NCI Office of Dietary Supplements, IUPAC, University of Hawaii, and others.	https://cabig.nci.nih. gov/inventory/ Infrastructure/

SILVER COMPATIBILITY

In addition, several tools were developed outside of caBIG™ and submitted for compliance review. The majority of developer projects in caBIG™ are developing to Silver level compatibility. *https://cabig.nci.nih.gov/guidelines_documentation/Silver_Review/#silver*

$^{SC(R)}$ = Tools currently under review for Silver Compatibility [This includes: Cancer Central Clinical Participant Registry (C3PR), caTRIP, Cancer Models Database (caMOD)]

^{SC} = Silver Compatible Products [This includes: caTissue Core, cancer Text Information Extraction System (caTIES), caIntegrator, SEED, Transcript Annotation Prioritizing and Screening System (TrAPSS), Clinical Data Exchange/Lab Integration Hub (caXchange), Function Express (caFE) Protein Information Resource (PIR), RProteomics, Proteomics Laboratory Information Management System (protLIMS), Reactome, Cancer Bioinformatic Infrastructure Objects (caBIO)]

BRONZE COMPATIBILITY

The caBIG™ Bronze certification program is a mechanism for software products not created as part of the caBIG™ program to be certified as compliant with the caBIG™ compatibility guidelines at the Bronze level. Current Bronze-level products include: BioXM, TrialCheck, Biological Specimen Inventory System (BSI) *https://cabig.nci.nih.gov/guidelines_documentation/bronze/#bronze*

APPENDIX B

Organizations Participating in the caBIG™ Initiative

Over the course of the three-year Pilot Phase of caBIG™, more than 1,000 individuals from nearly 200 organizations contributed time and expertise as part of the caBIG™ community. While it is not possible to recognize everyone individually, it is also impossible to overstate the centrality and importance of their efforts in the development of technology, policies, best practices, and other resources. Names and contributions of participants can be seen in meeting notes of the individual workspaces, which may be accessed through the caBIG™ Web site: https://cabig.nci.nih.gov/index_html/workspaces/index_html/. Individuals and teams whose outstanding contributions have merited a Recognition Award are listed at https://cabig.nci.nih.gov/events_folder/2006/2006_AnnualMeeting_Day_2/Recognition_Program_final.pdf/view/, and https://cabig.nci.nih.gov/News_Folder/caBIG_AwardeePDF.pdf/.

The following is a list of organizations participating at the conclusion of the Pilot Phase. The NCI-designated Cancer Centers that were the initial participants in the caBIG™ Pilot Phase are ***marked in bold****. Their pioneering activity set the stage for the growth that followed.*

Cancer Center/Medical Center Participants

- **Albert Einstein Cancer Center**
- **Arizona Cancer Center – University of Arizona**
- Baylor College of Medicine
- Brown University
- **The Burnham Institute**
- **Case Western Reserve University School of Medicine**
- **Cancer Research Center of Hawaii**
- **Chao Family Comprehensive Cancer Center – University of California at Irvine**
- **City of Hope National Medical Center & Beckman Research Institute**
- **Cold Spring Harbor Laboratory**
- College of William and Mary
- Columbus Children's Research Institute
- Drexel University
- **Duke Comprehensive Cancer Center**
- **Norris Cotton Cancer Center – Dartmouth-Hitchcock Medical Center**
- Dana-Farber Cancer Institute – Harvard University
- DPRN Coordinating Center
- Emory University School of Medicine
- **Fox Chase Cancer Center**
- **Fred Hutchinson Cancer Research Center**
- George Washington University
- Georgia Institute of Technology
- **Lombardi Comprehensive Cancer Center at Georgetown University**
- Group Health Cooperative
- **H. Lee Moffitt Cancer Center & Research Institute – University of South Florida**
- Helen F. Graham Cancer Center, Christiana Care Health Services
- **Herbert Irving Comprehensive Cancer Center – Columbia University**
- **Holden Comprehensive Cancer Center – University of Iowa**
- **Indiana University**
- **Jackson Laboratory**
- **Kimmel Cancer Center – Thomas Jefferson University**
- Louisiana Cancer Research Consortium
- Louisiana State University Health Sciences Center
- **Mayo Clinic Comprehensive Cancer Center**
- Massachusetts Institute of Technology Center for Cancer Research
- **MD Anderson Cancer Center – University of Texas**
- **Memorial Sloan-Kettering Cancer Center**
- Mouse Genome Informatics
- **New York University Cancer Institute**
- **Robert H. Lurie Comprehensive Cancer Center of Northwestern University**
- **The Ohio State University Research Foundation**
- **Oregon Health and Science University**
- **Penn State College of Medicine**
- **Meyer L. Prentis / Karmanos Comprehensive Cancer Center of Metropolitan Detroit**
- Roswell Park Cancer Institute
- **The Sidney Kimmel Comprehensive Cancer Center – Johns Hopkins University**
- Siouxland Hematology Oncology Associates

- **Siteman Cancer Center–Washington University School of Medicine**
- **St. Jude's Hospital**
- Stanford University
- Sunnybrook Health Sciences Centre
- Sylvester Comprehensive Cancer Center–University of Miami
- Texas Tech University
- **University of Alabama at Birmingham**
- University of Arkansas for Medical Sciences
- University of California at Berkeley
- **UC Davis Cancer Center – University of California, Davis**
- University of California, Los Angeles
- **University of California, San Francisco Comprehensive Cancer Center**
- Rebecca and John Moores UCSD Cancer Center – University of California, San Diego
- **University of Chicago Cancer Research Center**
- **University of Colorado Health Sciences Department of Preventive Medicine and Biometrics (UC-PMB)**
- University of Maryland
- **University of Michigan Comprehensive Cancer Center**
- **University of Minnesota Cancer Center**
- **University of Nebraska Medical Center – Eppley Cancer Center**
- University of New Mexico Cancer Research and Treatment Center
- **UNC Lineberger Comprehensive Cancer Center – University of North Carolina, Chapel Hill**
- **University of Pennsylvania – Abramson Cancer Center**
- **University of Pittsburgh Medical Center**
- University of the Sciences in Philadelphia
- **University of Southern California – Norris Comprehensive Cancer Center**
- University of Utah
- **University of Wisconsin Paul P. Carbone Comprehensive Cancer Center**
- **University of Vermont**
- **University of Virginia**
- **Vanderbilt-Ingram Cancer Center**
- **Virginia Commonwealth University – Massey Cancer Center**
- Virginia Polytechnic Institute and State University
- **Wake Forest Comprehensive Cancer Center**
- **The Wistar Institute**
- **Yale Cancer Center**

Government Organizations

- Argonne National Laboratory
- California Institute of Technology – Jet Propulsion Lab
- CDC National Center for Health Statistics
- Department of Veterans Affairs
- U.S. Food and Drug Administration
- Kentucky Cancer Registry
- Lawrence Berkeley National Laboratory
- National Library of Medicine
- NCI Cancer Diagnosis Program
- NCI Cancer Imaging Program
- NCI Cancer Therapy Evaluation Program
- NCI Consumer Advocates in Research and Related Activities (CARRA) or Patient Advocate Representative
- NCI Center for Bioinformatics
- NCI Center for Cancer Research
- NCI Division of Cancer Control and Population Sciences
- NCI Division of Cancer Prevention
- NCI Division of Cancer Treatment and Diagnosis
- NCI Division of Epidemiology and Genetics
- NCI Office of Biorepositories and Biospecimen Research
- NCI Office of Cancer Content Management
- NCI Office of Centers, Training, and Resources
- NCI Office of Communications
- NCI Office of the Director
- NCI Operational Research Office
- Virginia Cancer Registries

Cooperative Groups and Community Clinical Oncology Program Centers (CCOP)

- American College of Radiology Imaging Network
- Coalition of Cancer Groups
- Eastern Cooperative Oncology Group
- Cancer and Leukemia Group B (CALGB)
- Illinois Oncology Research Association CCOP
- National Surgical Adjuvant Breast and Bowel Project
- North Central Cancer Treatment Group
- Radiation Therapy Oncology Group
- Southwest Oncology Group
- Wichita CCOP

Consortiums & Non-Profit Organizations

- Clinical Data Interchange Standards Consortium, Inc. (CDISC)
- The Hastings Center
- Health Level 7 Seven, Inc. (HL7)
- Internet2
- Life Sciences Society (LSS)
- Oncology Nursing Society
- Quality Assurance Review Center (QARC)
- Translational Genomics Research Institute (TGen)

Industry Partners

- 3rd Millennium, Inc.
- 5AM Solutions
- 9Star Research, Inc.
- Advanced Clinical Software
- Agfa HealthCare
- Akaza Research
- Alpha-Gamma Technologies, Inc.
- Amgen Inc.
- Apelon, Inc.
- Aptia Systems, Inc.
- BioClinformatics LLC
- Business Technologies Group
- Capital Technology Information Services, Inc. (CTIS)
- Cedara Software Corporation
- Center for Health Research, Hawaii (CHRH) – Kaiser Permanente Hawaii
- Cerner Corporation
- Constella Group, LLC / Constella Health Sciences
- The Daedalus Group, Inc.
- Dataworks Development, Inc.
- Ekagra Software Technologies, Ltd.
- The EMMES Corporation
- First Clinical Research LLC
- First Genetic Trust
- General Electric Company / GE Global Research Center
- GulfStream Bioinformatics Corporation
- HK Stevenson, Inc.
- IBM Corporation
- IMPAC Medical Systems, Inc.
- INCOGEN, Inc.
- Independent Oncology Services
- Information Management Services, Inc. (IMS)
- Intel Americas Inc.
- Kitware, Inc.
- Mesh Ferguson
- Nortel Government Solutions
- OmniComm Systems, Inc.
- Optra Systems
- Oracle Corporation
- PercipEnz Technologies
- Persistent Systems Limited
- PSI International, Inc.
- QuaTeams
- Riverain Medical
- RTI International, Inc.
- Science Applications International Corporation (SAIC)
- ScenPro, Inc.
- Semantic Bits, LLC
- Siemens Corporate Research
- SRA International, Inc.
- Stone Bond Technologies
- TeraMedica Inc.
- TerpSys
- Terrapin Systems, LLC
- Theradex®
- Velos, Inc.
- Vivalog Technologies
- Vital Images, Inc.
- Westat
- Xcalibur

Pharmaceutical Industry Participants

- GlaxoSmithKline plc.
- Millennium Pharmaceuticals, Inc.
- Pfizer, Inc
- RadPharm

Attorneys

- Hall, Render, Killian, Heath & Lyman, P.C.

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NATIONAL INSTITUTES OF HEALTH

NATIONAL® CANCER INSTITUTE

NIH Publication No. 07-6244
Printed November 2007