# THE COMPUTERWORLD HONORS PROGRAM

## CASE STUDY

**ORGANIZATION:**

National Cancer Institute (NCI)

**PROJECT NAME:**

Cancer Biomedical Informatics Grid, or caBIG™

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

## Summary

The cancer Biomedical Informatics Grid, or caBIG™, is a voluntary virtual informatics infrastructure that connects data, research tools, scientists, and organizations to leverage their combined strengths and expertise in an open environment with common standards and shared tools. Effectively forming a World Wide Web of cancer research, caBIG™ promises to speed progress in all aspects of cancer research and care including etiologic research, prevention, early detection, and treatment by breaking down technical and collaborative barriers.

Researchers in all disciplines have struggled with the integration of biomedical informatics tools and data; the caBIG™ program demonstrates this important capability in the well-defined and critical area of cancer research, by planning for, developing, and deploying technologies which have wide applicability outside the cancer community.  Built on the principles of open source, open access, open development, and federation, caBIG™ infrastructure and tools are  open and readily available to all who could benefit from the information accessible through its shared environment.

caBIG™ is implemented by the cancer and biomedical research community, in collaboration with nonprofit and industry partners (non cancer-specific organizations), with coordination and oversight provided by the National Cancer Institute Center for Bioinformatics (NCICB). As part of the activities involved in building the Grid, participating NCI-designated Cancer Centers and industry partners are developing or providing standards-based biomedical research applications, infrastructure, and data sets. The implementation of common standards and a unifying architecture ensures interoperability of tools, facilitating collaboration, data sharing, and streamlining research activities across organizations and disciplines.

Development of technical infrastructure and collaborative community is never easy, but as the transforming power of an integrated infrastructure becomes more widely appreciated, an ever increasing number of stakeholders are examining their own environments and asking how they can build interconnecting links through caBIG™. The foundational components of caBIG™ are readily available as building blocks, connectors, and tools, because much of the information and processes associated with cancer prevention, care, and research share fundamental elements or approaches with other health challenges. Moreover, sharing and integrating functional genomics

ORGANIZATION:
*National Cancer Institute (NCI)*

PROJECT NAME:
*Cancer Biomedical Informatics Grid, or caBIG™*

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

and clinical trial data can improve cancer prevention and treatment beyond any one country; the burden of cancer is international. caBIG™ is collaborating with the National Cancer Research Institute in the United Kingdom to enable a richer analysis of complex relationships between, for example, patterns of gene expression and prognosis or response to treatment.

## Introductory Overview

Launched in February 2004, caBIG™ was designed and developed in collaboration with 50 NCI-designated Cancer Centers and over 30 other organizations. Over 800 individuals now contribute to the caBIG™ initiative. caBIG™ activities are organized in workspaces; areas of focus that are developing applications, infrastructure, standards, policy documents, and other resources. There are currently nine workspaces: Architecture, Clinical Trials Management Systems, Data Sharing and Intellectual Capital, Integrative Cancer Research, In Vivo Imaging, Strategic Planning, Tissue Banks and Pathology Tools, Training, and Vocabularies and Common Data Elements. Collectively, caBIG™'s Workspaces are not only building the foundation for caBIG™, they are also driving caBIG™'s goals, priorities and activities. Voluntary participation is welcome and encouraged, and will ultimately ensure caBIG™'s long-term success.

caBIG™ is already delivering demonstrable cancer and biomedical research products. In its first two years, caBIG™ has launched over 90 individual projects including: the first iteration of the caBIG™ Compatibility Guidelines; end-to end solutions like caARRAY and genePattern, that provide micorarray tools at both ends of the process; caWorkbench, providing analysis capabilities for molecular pathways; caTIES, caTISSUE CORE, and Clinical Annotation Engine, a set of tissue banking tools that can be used to track and mine tissue samples; and the Cancer Central Clinical Data suite of clinical trials management tools for managing clinical research data across sites and time, including key functions like adverse event reporting. The testbed release of the grid architecture, dubbed caGrid, is available and version 1.0 will be released in the Fall.

During the 2006 calendar year, 40-plus new products are expected to be delivered, to include biomedical tools and datasets, as well as white papers, policies, guidelines, and training materials. All products produced by caBIG™ are openly available for use by the caBIG™ community and beyond. caBIG™ products developed by the Workspaces are also increasingly available online. (Detailed descriptions of these tools are available in the accompanying caBIG™ DVD).

NCI recognizes that ultimately, the evolution of the caBIG™ network should be accompanied by the growth of a self-sustaining caBIG™ community. Within the cancer community, caBIG™ started with NCI-designated Cancer Centers and is now reaching out to NCI's Specialized Programs of Research Excellence (SPOREs), which promote interdisciplinary research among the basic and clinical sciences; NCI's Clinical Trials Cooperative Group Program, that involves researchers, cancer centers, and community physicians; and other NCI programs, and the broader cancer community. Patient advocates have played a critical role in caBIG™ from the beginning.

Equally important, caBIG™ participants include industry partners (e.g., Pharmaceutical, Biotechnology, IT), other National Institutes of Health (NIH) researchers, clinicians, and informaticians, U.S. federal agencies, and international partners. Where possible, caBIG™ is collaborating and coordinating with other health and biomedical IT initiatives and activities (public and private) to further the goal of creating an interconnected network to support the vision for overcoming cancer. Specific partnerships between caBIG™ and other NIH components, Federal agencies and international initiatives are also being discussed. All of these groups share a com-

ORGANIZATION:
*National Cancer Institute (NCI)*

PROJECT NAME:
*Cancer Biomedical Informatics Grid, or caBIG™*

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

mon commitment to the importance of open and shared biomedical informatics tools, standards, infrastructure and data.

Our DVD that accompanies this submission provides much more detail on the caBIG™ program including FAQs, lists of tools, articles, program updates, participant listings, and workspace contact information. The DVD also allows users to navigate to a "caBIG™ toolkit" representing the current status of the caBIG™ initiative's software development efforts to create interoperable and standards-based biomedical informatics tools.

## Benefits

To me the biggest thing about caBIGTM is how it brought everyone in the community together. Getting all of the agencies (NCI, FDA), the Cancer Centers and all the various bodies to communicate and to start talking about what they have been doing so that they can get on the same page will foster a breakthrough. Even if we were never to even get a single software product out, the standards that are developed by the program would make a significant difference in the war on cancer.

- Diane Paul, Patient Advocate

In 2005 it was estimated that 1.4 million people received the horrible news that they had cancer. And in 2005, an estimated 570,280 people died from cancer. The caBIG™ program is working to provide real solutions to help reduce the number of patients that are dying or suffering from cancer. caBIG™ will involve the entire cancer research community, including those who conduct basic science research on the origins and mechanisms of cancer, those who study prevention, early detection, and treatment, and those who work on clinical trials to bring effective new diagnostics and treatments to patients. Through caBIG™, basic scientists will be better able to integrate disparate forms of data from their own laboratory, as well as from other research laboratories across the world. They will be able to integrate information based on tumor pathology; data from RNA, DNA, and protein expression levels (integrating genomic, expression array, and proteomic experiments); and data collected from patients involved in clinical trials. caBIG™ will increase the strength and scope of the experiments done in each participating center, thereby generating broader and more meaningful conclusions that can be translated more rapidly into better patient outcomes. caBIG™ will also help to enable another important step—taking the most promising ideas from bench to bedside, and back to the bench again.

These activities are the result of not only the cooperative development of new software and informatics infrastructure, but also in the creation of an environment supporting collaboration among the broad range of participants. It is within the context of this diverse community that the caBIG™ project tools have been developed. By working closely together with other researchers on related classes of information, scientists and developers can ensure that their data is collected and stored in a manner that enables it to be shared, queried, retrieved and integrated with that from others. This allows data collected by the community to be combined and used for purposes unexpected by the original collectors of the material. It is these integrative and translational activities which will provide the best use of the original data, and provide the most novel and important scientific insights.

Although caBIG™'s most important accomplishments and contributions still lie in the future, it has the power to redefine how cancer research is conducted and shared. caBIG™ is hastening the

ORGANIZATION:
*National Cancer Institute (NCI)*

PROJECT NAME:
*Cancer Biomedical Informatics Grid, or caBIG™*

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

time when patients live with, rather than die from cancer.

## The Importance of Technology

The caBIGTM program differs from other grid efforts in that it represents a strongly typed grid with several layers of metadata and a sophisticated hierarchical metadata management system. Cancer research is an ideal application area to motivate important advances in grid technology because of the heterogeneous nature of biomedical data and the acknowledged need to coordinate and share resources across that community.

– Joel Saltz M.D. Ph.D., Professor and Chair of Biomedical Informatics, Arthur G. James Cancer Hospital & Richard J. Solove Research Institute, Ohio State University

The cancer research community is in the midst of an explosion of knowledge about cancer as a disease process – beginning to understand cancer not by what can be seen or touched – 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. As scientists understand more completely the steps of the cancer process, they will identify the specific molecular targets in that process that are vulnerable to preemption. This cannot be achieved without greater interconnectivity and coordination across the cancer enterprise. This requires seeing cancer as a systems problem that will require a systems solution. Although cancer is being unraveled rapidly at the genomic and proteomic levels, researchers have not concomitantly developed the seamless system needed to capitalize on discoveries. To universally integrate personalized medicine into cancer prevention, diagnosis and treatment, researchers and clinicians must be able to gain rapid access to multiple types of specific information about an individual patient -- information to which they do not currently have easy access. A new generation of medicine will require incorporation of shared information technologies.

The information infrastructure revolution that has transformed business has had slow uptake in biology and medicine. Within the research community there exists a "Tower of Babel" problem. Research teams cannot easily understand data collected by, or share data with, other medical research teams working on the very same disease. Efficient, effective collaborations are blocked by these "language" and data sharing problems. Scientists have a difficult time integrating the various types of data they collect in a manner that will allow them to ask and answer important questions about how a disease works, and what they can do to stop it. Medical research teams have operated, in effect, as cottage industries, each collecting and interpreting data using a unique language of their own making and in virtual isolation from other teams. Biomedical informatics has the potential to be the powerful critical means to achieve the necessary degree of integration as it provides the mechanisms and tools to support standardized sharing, management and analysis of diverse data across the bench-to-bedside continuum and back.

By making use of standardized vocabularies and objects, defined by community interaction, and reflecting the needs and requirements of those participants within the community, data under the caBIG™ project is collected in a previously-agreed upon structure, using shared and standardized vocabularies, and accessed via a standard software mechanism. It is with these capabilities along with well-defined security and privacy tools and processes that the data can be accessed, shared, and ultimately integrated. With integrated data and standard protocols, user-friendly desktop tools have been (and are continuing to be) developed to allow end users to seamlessly query data from multiple institutions in many disciplines. With the addition of

ORGANIZATION:
*National Cancer Institute (NCI)*

PROJECT NAME:
*Cancer Biomedical Informatics Grid, or caBIG™*

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

caBIG™ standardized support, documentation, and training using user-friendly desktop tools, a growing range of software is being developed by the program.

Informatics at the Core of the caBIGTM Program

Recognizing the transformational power of an interoperable biomedical informatics infrastructure to overcome obstacles, caBIG™ is constructed around three interrelated areas of informatics:

•Bioinformatics provides cancer and biomedical researchers with tools, infrastructure and analytic methodologies necessary to manage and harvest insights from the large volumes of data generated by novel types of research such as molecular biology, genomics and proteomics.

•Medical or clinical informatics enables the management, analysis and dissemination of clinical and public health data, and includes the use of informatics infrastructure and applications such as clinical trial management systems, electronic health records, and cancer registries.

•Biomedical informatics, an innovative synergy between bioinformatics and clinical informatics, offers infrastructure, tools, techniques and applications that bridge the two other areas and creates a mechanism that facilitates the sharing of data along the continuum from research bench to clinical bedside and back. It offers the prospect of integrating individual patient data from clinical care into the clinical research environment, and back into clinical care or basic science research.

caBIG™ is being built on open source, open access, open development, and federation principles. Anyone can gain access to caBIG™ software (and its component parts) at no cost, modify it to suit his or her needs, and contribute to its ongoing development. Many organizations are working together to realize its full potential, including for-profit companies that wish to "add value" to caBIG™ – i.e., enhance caBIG™'s function or ease of use – and lease or sell those improvements to interested caBIG™ users. As an open source model, source code for all program-funded caBIG™ software tools is available to end-users. This approach is consistent with the philosophy supporting development of the "knowledge commons," which is created and fed by free and open distribution of intellectual property on the internet.

## Originality

For the first time, an informatics system stands to meld the contextual knowledge that has been developed in the biosciences. caBIGTM is forward-looking in the same way that the internet infrastructure was, and that has effectively served the community for many years.

-Frank Manion, Chief Technology Officer

Fox Chase Cancer Center, Philadelphia PA

caBIG™ is a distinctive and ambitious undertaking, as no known precedent exists for a bioinformatics engineering initiative of this scale. The NCI is helping caBIG™ become "the World Wide Web of cancer research." Researchers from around the world are gaining open access to the common platform of caBIG™, to be able to use common tools, and rapidly convert, relate, and analyze data from different sources.  Members of the research community also actively contribute to caBIG™ activities based on their needs and interests, as the  work of many groups across the government, academic, and private sectors is crucial to the success of caBIG™.

caBIG™ has already started to inspire scientists throughout the cancer research community and

ORGANIZATION:
*National Cancer Institute (NCI)*

PROJECT NAME:
*Cancer Biomedical Informatics Grid, or caBIG™*

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

beyond. Bioinformatics researchers have required the implementation of standards since the field first started to expand during the explosion of DNA sequencing data when the genome project began. It was the need for a consistent means to represent DNA and protein sequence data, coupled with a growing range of associated annotation data, that has driven the caBIG™ project to develop a set of standards that can be used to integrate data and connect applications. This process is unique in that it does not require the extra time and cost of additional software activities to standardize and integrate the applications. The community has collaborated closely to incorporate this standardization and integration right into the process. The tools and applications are out to the community cheaper and quicker.

## Success

caBIG™ opens up a ton of opportunities to do research on unusual forms of cancer. This is particularly important where there are not enough people in any given region for a statistically valid study. When you can network a bunch of cancer centers together, you can assemble lots of different cases, and with that integrated information, benefit whole new cancer populations.

Virginia Hetrick, Patient Advocate

Professor, DeVry University

The progress of the caBIG™ program in a few short years has been incredible: from the development of a strong and vibrant community, through the selection and implementation of key infrastructure technologies, to the ongoing development and deployment of a suite of tools now utilized in cancer research settings across the country and beyond. caBIG™'s community has grown to more than 800 active participants, working in close cooperation in a range of domain areas and contributing to the shared development and adoption of data models and common data elements to represent their informatics work. Such harmonization of the diverse subject areas represented by project participants is integral to the development of caBIG™'s underlying shared infrastructure.

caBIG™'s foundational program has facilitated the development of tools which can now be deployed in scientific research laboratories to enable the collection, analysis and visualization of data, as well as providing a host of mechanisms for collaboration between members. There are more than 50 working products in the caBIG™ program which are available now, with many others coming online in the near future. Similarly, there are already a wide range of data services and sources being supplied from laboratories throughout the country using caBIG™ compatible interfaces. As the caGrid technology moves from its current 0.5 release to the caGrid 1.0 release, information from many more data sources will become available and queryable through simple and straightforward user interfaces currently under development.

Subsequently, caBIG™ has seen a steadily growing range of participants on the Grid bringing an ever-widening knowledge base. The internet, much like caBIG™, had a small beginning— once just two computers on the network—and the internet has grown through open and interoperable systems, to become a critical, worldwide resource. The caBIG™ project is taking a very similar open and community-driven approach which is likely to demonstrate similar success.

The community has already responded, creating thousands of new data elements and vocabulary concepts that can be re-used, and all applications developed under caBIG™ are at the Silver level of maturity, which requires that all relevant software data elements, vocabularies can be re-used

ORGANIZATION:
*National Cancer Institute (NCI)*

PROJECT NAME:
*Cancer Biomedical Informatics Grid, or caBIG™*

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

and harmonized and that the APIs are exposed in such a way that they can integrate with the under-development caGrid. Several early adopters within the community have also piloted their software on caGrid 0.5. By exposing their software API on caGrid 0.5, these early adopters of caGrid 0.5 have provided the cancer research community access to a protein information resource (PIR, Georgetown University), a protein mass spec analytical service (RProteomics, Duke), a microarray database (caArray, NCICB and Georgetown University) and a pathology report text extraction app (caTIES, UPMC). Not only has this effort provided tools to the cancer research community but has served as a means of developing reference implementations to guide later caGrid adoption throughout the program. As a result, there are already working caGrid nodes throughout the country, with more being added all the time.

caBIG's™ progress and developments are designed to be extensible into a wide range of disciplines. Researchers in disease areas other than cancer are already utilizing caBIG's™ infrastructure and adopting its approach to meet their research needs.

## Difficulty

Since nothing of this kind or scale had ever been successfully attempted, merely getting the stakeholder community educated and involved in the effort was a challenge. In order to initially engage the cancer community, all of whom are located at the more than 60 NCI-designated Cancer Centers spread throughout the country, it was necessary to visit the centers and meet the participants. In order to meet this challenge, the NCICB launched the preparatory phase of caBIG™ in July 2003 by engaging designated NCI Cancer Center staff in informational seminars held on each coast. Over 100 individuals participated in these discussions. The purpose of these discussions was to inform the cancer centers about the caBIG™ initiative goals, objectives, and timelines. Careful consideration was given to create a message designed not only to inform, but to generate enthusiasm for the initiative. Immediately following these sessions, five teams conducted onsite cooperative development meetings where Cancer Centers discussed their informatics based strengths, needs and potential contributions in greater detail. caBIG™ program staff visited 49 Cancer Centers in 42 days. The combined expertise of the visiting teams allowed the caBIG™ program team to effectively communicate on a peer-to-peer level with the local Cancer Center experts. The visits resulted in a detailed accounting and prioritization of biomedical informatics related issues, interests and capabilities that the caBIG™ program sought to address.

The meetings with Cancer Centers were critical towards the development of a program that truly engaged the stakeholders, and significantly moved the program towards self-sufficiency. These meetings gave the Cancer Center bench researchers, clinical trialists and informatics specialists all a chance to contribute their ideas, needs and capabilities to the developing program during its early formative stages. The structure of the program itself, and its early emphasis came directly from the participating stakeholders. This has not only led to a vibrant program, it has also led to a program that the stakeholders feel a strong sense of ownership for. caBIG™ is overall driven by and for the community.

The challenges of developing systems which are not only capable of sharing data in a common format, but are also able to integrate that data across sites, laboratories, and subject areas, is the chief technical difficulty to be overcome by the caBIG™ program. The critical first steps were to develop an architecture and software infrastructure which allows the effective sharing of infor-

ORGANIZATION:
*National Cancer Institute (NCI)*

PROJECT NAME:
*Cancer Biomedical Informatics Grid, or caBIG™*

LOCATION:
*Rockville, Maryland, United States*

YEAR:
*2006*

STATUS:
*Laureate*

CATEGORY:
*Science*

NOMINATING COMPANY:
*Booz Allen Hamilton Inc.*

mation, and to develop and use software tools and processes to allow the members of the community to re-use and harmonize the data models and vocabularies that they use to describe their data. By taking advantage of innovative tools like the caDSR (cancer data standards repository) the EVS (enterprise vocabulary system) and the caGrid, a geographically dispersed community of researchers and informaticians can query data in a uniform and integrated fashion from tools placed right on their desktops. The caBIG™ program is leveraging existing software standards and open platforms wherever possible, and closely working with experts in the community, to develop what is necessary to achieve the goals of the program.

Beyond the difficulty of coordinating the development of a suite of interoperable biomedical informatics tools, and the creation of shared vocabularies and data elements, was the challenge of creating a self-sustaining community of experts to shepherd and lead the development of the program. It is only with a committed and coordinated group of developers that a program as large in scope and as aggressive in timeline as caBIG™ can be created. Although the field of cancer research is in dire need of better means of data integration and collaboration, there has not been a history of software standards within the bioinformatics community. Addressing this dual challenge of complex software and infrastructure development requirements and the molding of a community to support them has been the critical achievement of caBIG™.

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.

caBIG™ has been supported by the National Cancer Institute from its inception as a critical element in the Institute's strategic plans and a key enabler of its vision to eliminate suffering and death due to cancer by 2015.