caBIG™ for Researchers

What is caBIG™?

Today's researchers are faced with an overwhelming volume of valuable data. Data are generated by a variety of sources and using data from a given source requires understanding the terminology used to describe that data and reformatting the data in a way that is useful.

The vision of caBIG™ is a world in which biomedical data and applications are seamlessly integrated such that they can be used collaboratively to advance cancer research and ultimately affect cancer treatment and prevention.

caBIG™, which stands for cancer Biomedical Informatics Grid, is an ongoing program of the National Cancer Institute. The goal of the caBIG™ program is to create a so-called "World Wide Web" of cancer research data and tools. Such a network will have a profound influence on how research is conducted, how care is provided and how the cancer community interacts with the biomedical research enterprise.

What approach is being taken to achieve caBIG™ program goals?

The caBIG™ program is connecting researchers through a sharable and interoperable software infrastructure. In caBIG™, interoperability is defined as the ability to exchange information and to use the information that has been exchanged.

Interfaces

Interoperability is defined at the interfaces between information systems – interfaces are how data get in and out of these systems. Of course, the term "interface" can also be used to refer to a graphical user interface which is how end-users typically interact with software. In caBIG™, the focus is on *programmatic* interfaces – something an end-user isn't typically exposed to directly. The rules and infrastructure being developed through caBIG™ apply to these programmatic interfaces; a general term used to describe compliance with these rules is "caBIG™ Compatibility." Thus, the goal of caBIG™ Compatibility is the creation of interoperable interfaces that enable the sharing of scientific content.

Analogy of a City

An analogy for the importance of standard interfaces is that of a city. In cities, architects are free to design buildings that perform many different functions and that take many different forms. For a city to function, buildings must conform to certain specifications to receive electricity, water, steam, mail, etc. For most of us, these functions happen behind the scenes in our everyday lives. Likewise, the standardized programmatic interfaces that support interoperability in caBIG™ are, for the most part, hidden to the end user.

What does this all mean to researchers?

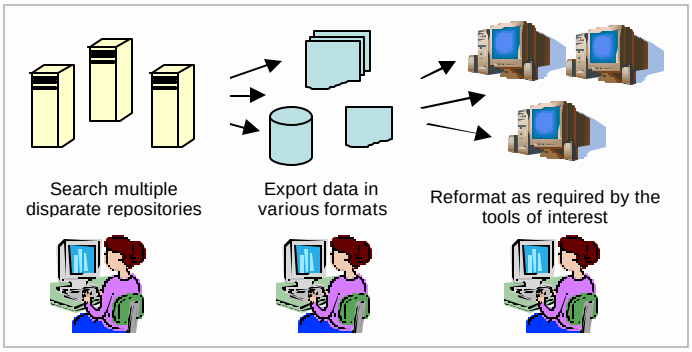
Most importantly, caBIG™ can benefit your research. Consider the following scenarios:

Scenario 1: Data aggregation and analysis workflows

I would like to form an integrative analysis of mouse microarray expression data involving treatment with rapamycin.

Using systems that are not caBIG™-compatible, a researcher has to:

- Search all known repositories for relevant data (data may or may not be appropriately annotated with the compound treatment)
- Export the data in whatever format that repository provides
- Reformat the data to comply with the format required by the tools of interest. It is likely that each tool will require a different format.

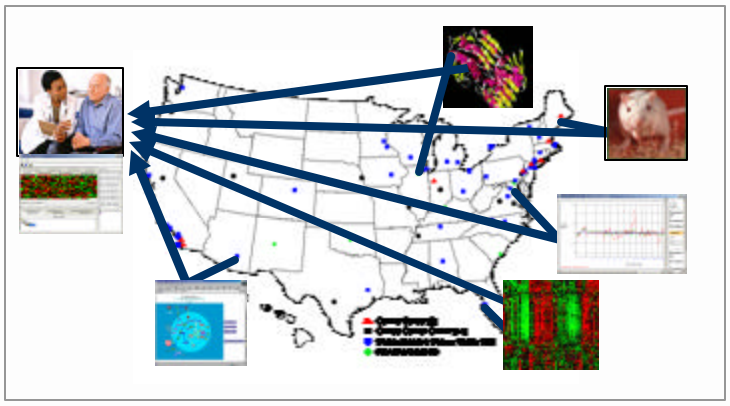


Using caBIG™ compatible systems, the user can query for microarray repositories that have data from samples treated with rapamycin. The researcher could then select to have this data automatically imported into the analysis tool or sent to a remotely available analysis service for processing. Because both the microarray data repository and the analysis tools/services use the same format and terminology to exchange data, this process can occur seamlessly.

Scenario 2: Data integration

What are all the available tissue specimens from HER2/neu-positive patients who were diagnosed with a lobular carcinoma and are BRCA1 positive?

This scenario represents a use case that was addressed in caBIG™ through the caTRIP project – the cancer Translational Research Informatics Platform. caTRIP is software that enables users to build complex queries for those instances when data are stored across multiple caBIG™-compatible systems.



In this example, the data needed to answer this question were distributed across four information systems. In the non-caBIG™ environment, a researcher would have to search each system independently to get the information she or he needs. Such a search would involve learning how to use four different systems and determining how to merge the results.

Through the caBIG™ program, caBIG™-compatible interfaces were developed for the four relevant information systems. Using caTRIP, the researcher can now ask such questions using a single piece of software that is able to integrate data from multiple systems.

Frequently Asked Questions

Q: Do caBIG™ software systems look any different than the systems I'm familiar with?
A: Not really. In many cases, pre-existing systems have been adapted to be caBIG™ compatible.

Q: Will I have to use different terminology than I'm used to?
A: Not necessarily. In many cases, mapping of terms happens behind the scenes. As with any software, this will be case-by-case.

Q: Do I have to install the software locally? Or are these centralized resources?
A: In some cases, software is installed locally; in other cases, the systems are completely web-based.

Q: If I participate in caBIG™, do I have to make all of my data public?
A: No. The software infrastructure being built in caBIG™ gives scientists intricate control how and when their data are shared. **Data can be kept private; shared with defined collaborators; or, made public as deemed appropriate.**

Q: I have some microarray data that are stored in a non-caBIG™ database at my institution. I would like to analyze the data with caBIG™ analysis tools. How do I do this?
A: There are a few of options here:

a) You can load your data into caArray, a caBIG™-compatible microarray data repository. A number of the caBIG™ microarray analysis tools, including GenePattern and the geWorkbench analysis packages, can read data in directly from caArray. These packages are easily installable on your local computer. **It is important to note that data in caArray can be kept private so that loading and storing your data does not expose it publicly before you are ready to do so.**
b) Data can also be imported into these tools using non-caBIG™ approaches (as mentioned above, most were developed prior to caBIG™). In most cases, different tools will require different formats; therefore, multiple reformatting steps may be required.
c) Your organization may be interested in building a caBIG™-compliant interface for your institution's existing database. This is, of course, a longer-term strategy.

Q: *How is the Mouse Models of Human Cancers Consortium involved in caBIG™?*
A: The MMHCC partnered with the NCI Center for Bioinformatics from its inception to deploy the *e*MICE website, and has collaborated with the caBIG™ program since it was established in 2003. There are several caBIG™ resources whose development is primarily driven by concepts and input from MMHCC members. The Cancer Model Database (caMOD) is a caBIG™ compliant repository of information about animal models of human cancers that is a resource for the public research community. In addition to information about mouse cancer models, caMOD can accommodate information about rat models; the ability to store information about zebrafish models is being added, and other species are under consideration. The Electronic Laboratory Management and Information Repository (caELMIR) is a cancer information system for managing experimental data for pre-clinical trials utilizing mouse models. caELMIR is an important asset for future bioinformatics integration of experimental therapeutics data from a variety of model systems with the outcomes of early-stage clinical trials.

Q: Is there training available to learn how to use the software? Who do I go to if I have questions?
A: The >40 software products of caBIG™ are at various levels of maturity. Some have dedicated, formal training support; others have detailed, self-guided training. The end-user support resources for each software product are listed at https://cabig.nci.nih.gov/tools/. For more information about the caBIG™ program, contact Mervi Heiskanen, heiskame@mail.nci.nih.gov, 301-451-6369, or visit the "About caBIG™" resources on the caBIG™ website at https://cabig.nci.nih.gov/overview.

A Selection of caBIG™ Resources for Researchers

caELMIR

What is it?	The **E**lectronic **L**aboratory **M**anagement and **I**nformation **R**epository (caELMIR) is a cancer information system for managing experimental data for pre-clinical trials utilizing mouse models.
Where can I find it?	https://gforge.nci.nih.gov/projects/elmer/ http://caelmir.compmed.ucdavis.edu:8080/caelmir/
How can I access it?	Request a user account on the website; your account information will be emailed to you.
How will it help me in my work?	This system connects to your colony management system and allows you to create studies and assign individual animals to cohorts within the study. caELMIR will support the collection of various data types for each animal such as demographic data and pathology information.

CPAS

What is it?	The open-source Computational Proteomics Analysis System (CPAS) contains an entire data analysis and management pipeline for Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) proteomics, including experiment annotation, protein database searching and sequence management, and mining LC-MS/MS peptide and protein identifications.
Where can I find it?	http://cpas.nci.nih.gov
How can I access it?	A user account is not needed to access the data in CPAS.
How will it help me in my work?	The public data stored at this site include data from the NCI Mouse Proteomic Technologies Initiative (MPTI), a component of NCI's Clinical Proteomic Technologies Initiative for Cancer (CPTI). CPAS architecture and features, such as a general experiment annotation component, installation software, and data security management, make it useful for collaborative projects across geographical locations and for proteomics laboratories without substantial computational support.

NCIA

What is it?	The National Cancer Image Archive (NCIA) is a searchable repository of in vivo cancer images that provides the cancer research community, industry, and academia with access to image archives to be used in the development and validation of analytical software tools
Where can I find it?	http://imaging.nci.nih.gov/ or http://ncia.nci.nih.gov
How can I access it?	Create a user account directly on the website.
How will it help me in my work?	NCIA provides access to imaging resources that will improve the use of imaging in today's cancer research and practice by increasing the efficiency and reproducibility of imaging cancer detection and diagnosis, and ultimately enabling the development of imaging resources that will lead to improved clinical decision support.

caMOD

What is it?	The cancer models database (caMOD) is a web-based resource that provides information about rodent and other animal models for human cancer to the public research community.
Where can I find it?	http://cancermodels.nci.nih.gov
How can I access it?	A user account is only needed for submitting data. Register on the website; account information will be emailed to you.
How will it help me in my work?	Users can retrieve information about the making of models, their genetic descriptions, histopathology, derived cell lines, associated images, carcinogenic interventions, microarray data, and therapeutic trials in which the models were used.

caArray

What is it?	caArray is a MIAME-compliant repository for the storage and retrieval of microarray data and associated annotations. caArray provides a MAGE-compliant programming interface to allow integration with a wide variety of tools for data analysis.
Where can I find it?	http://caarraydb.nci.nih.gov/caarray/
How can I access it?	A user account is only needed for submitting data or querying protected data. Register on the website; account information will be emailed to you.
How will it help me in my work?	The caArray data portal supports the submission of MIAME 1.1 level annotations and microarray data via web-based submission forms, as well as query and retrieval of these data.

caIMAGE

What is it?	The cancer Images Database (caIMAGE) is a submission and retrieval system for histology images.
Where can I find it?	http://cancerimages.nci.nih.gov
How can I access it?	A user account is only needed for submitting images. Create a user account directly on the website.
How will it help me in my work?	caIMAGE can be utilized to store and share histopathology images.

The caBIG™ Compatibility Guidelines and the caGrid infrastructure greatly facilitate interoperability among information systems. Integration is the primary focus for the current development efforts.

Connecting multiple applications will facilitate several types of analyses:

- Comparing outcomes for the same drug in different models
- Comparing outcomes of different protocols for a specific drug in a particular animal model
- Comparing outcomes after the use of different drugs in the same animal model
- Comparing across model species
- Comparing model and human data

For research on experimental therapeutics in model organisms, we are beginning to integrate the caMOD and caELMIR applications.

In a first step, we will expand the search capabilities of caMOD to retrieve data from caELMIR and link them to the page that lists therapeutic experiments in which a particular model was used. From there, users can retrieve information about individual animals used in the study, outcome data, and various other data stored in caELMIR. **Only caELMIR data that are public will be accessible in caMOD.**