

CAARRAY 2.2

User's Guide



Center for Bioinformatics

Revised December 23, 2008

CREDITS AND RESOURCES

caArray Development and Management Teams			
Development	Quality Assurance	Documentation	Project and Product Management
Bill Mason ²	Tom Boal ⁴	Rashmi Srinivasa ²	Maki Duncan ²
Dan Kokotov ²	Xiaopeng Bian ¹	Dan Kokotov ²	Rashmi Srinivasa ²
Rashmi Srinivasa ²	Leonie Misquitta ⁴	Doug Harley ²	Anand Basu ¹
Steve Lustbader ²	Arathi Reddy ⁵	Jill Hadfield ¹	Juli Klemm ¹
Max Shestopalov ²			
Doug Harley ²			
Systems and Application Support		Training	
Yeon Choi ³		Don Swan ³	
Andrea Johnson ³			
Bob Wysong ³			
¹ National Cancer Institute Center for Bioinformatics (NCICB)		² 5AM Solutions	³ Terrapin Systems
⁴ NARTech	⁵ Northern Taiga Ventures, Inc.		

Contacts and Support	
NCICB Application Support	http://ncicbsupport.nci.nih.gov/sw/ Telephone: 301-451-4384 Toll free: 888-478-4423

TABLE OF CONTENTS

Credits and Resources	i
Using the caArray User's Guide	1
Introduction to the caArray User's Guide	1
Organization of this Guide	1
User's Guide Text Conventions	2
Chapter 1	
About caArray	5
caArray Overview	5
Relationship of caArray to caBIG®	6
Chapter 2	
Getting Started in caArray	7
caArray Fundamentals	7
Requesting a User Account	10
Using caArray Online Help	13
Navigating the caArray User Interface	14
Chapter 3	
Navigating and Searching caArray	19
Browsing the caArray Repository	19
Searching the caArray Repository	22
Chapter 4	
Creating and Managing Experiments	29
Overview of an Experiment	29
Creating an Experiment	30
Updating An Experiment Proposal	56
Experiment Visibility	56

Chapter 5	
Curation Tools	63
Curation Tasks	63
Chapter 6	
Submitting Data to an Experiment	77
Managing Data	77
Downloading Files	92
Chapter 7	
Extracting Data from caArray	93
Downloading Data from caArray	93
Grid Availability	95
Extracting Data Programmatically by API	95
Chapter 8	
User Account Management	97
Roles in caArray	97
Managing User Accounts	98
Managing Collaboration Groups	99
Appendix A	
MAGE-TAB in caArray	103
caArray-Specific Handling of MAGE-TAB	103
Best Practices and Tips	105
Limitations of Annotation Data	108
caArray Validation of MAGE-TAB	109
Examples of MAGE-TAB documents	115
Appendix B	
caArray References	117
Index	119

USING THE CAARRAY USER'S GUIDE

This chapter introduces you to the *caArray 2.2 User's Guide* and suggests ways you can maximize its use.

Topics in this chapter include:

- [Introduction to the caArray User's Guide](#) on this page
- [Organization of this Guide](#) on this page
- *User's Guide Text Conventions* on page 2

Introduction to the caArray User's Guide

The *caArray 2.2 User's Guide* is the companion documentation to the caArray software application. The user's guide includes information and instructions for the end user about using caArray.

Organization of this Guide

The *caArray 2.2 User's Guide* contains the following chapters and appendices:

Using the caArray User's Guide—This chapter introduces you to the *caArray 2.2 User's Guide* and suggests ways you can maximize its use.

Chapter 1 About caArray—This chapter provides a caArray overview as well as a description of the relationship of caArray to caBIG™.

Chapter 2 Getting Started in caArray—This chapter provides details about launching caArray, navigating the interface, accessing online help and other links in the interface.

Chapter 3 Navigating and Searching in caArray—This chapter describes the process for creating and working with Contacts, namely individual and group contacts in caArray.

Chapter 5 Curation Tools—This chapter describes curation tasks available to all logged in users that relate to array designs, protocols and vocabulary terms.

Chapter 6 Creating and Managing Experiments—This chapter details instructions for creating and working with Experiments in caArray.

Chapter 7 Submitting Data to an Experiment—This chapter describes the processes for uploading, validating and importing array content and annotation data files into a caArray Experiment.

Chapter 8 Extracting Data from caArray—This chapter describes the processes for downloading data from the caArray repository.

Chapter 9 User Account Management—This chapter describes the process for creating user accounts and collaboration group accounts and managing the group accounts of caArray.

Appendix A MAGE-TAB in caArray—This appendix describes how MAGE-TAB documents are parsed, validated and imported into caArray. It also provides examples of the types of MAGE-TAB documents that are expected by caArray

Appendix B References—This appendix includes descriptions and links to references closely related to caArray technology and bioscience.

Index—This section of the guide provides a complete index.

User's Guide Text Conventions

Table 2.1 illustrates how text conventions are represented in this guide. The various typefaces differentiate between regular text and menu commands, keyboard keys, toolbar buttons, dialog box options and text that you type.


Convention	Description	Example
Bold & Capitalized Command Capitalized command > Capitalized command	Indicates a Menu command Indicates Sequential Menu commands	Admin > Refresh
TEXT IN SMALL CAPS	Keyboard key that you press	Press ENTER
TEXT IN SMALL CAPS + TEXT IN SMALL CAPS	Keyboard keys that you press simultaneously	Press SHIFT + CTRL and then release both.
Monospace type	Used for filenames, directory names, commands, file listings, and anything that would appear in a Java program, such as methods, variables, and classes.	URL_definition ::= url_string
Icon	A toolbar button that you click	Click the Paste button () to paste the copied text.
Boldface type	Options that you select in dialog boxes or drop-down menus. Buttons or icons that you click.	In the Open dialog box, select the file and click the Open button.
<i>Italics</i>	Used to reference other documents, sections, figures, and tables.	<i>caCORE Software Development Kit 1.0 Programmer's Guide</i>

Table 2.1 caArray Guide Text Conventions

Convention	Description	Example
<i>Italic boldface monospace type</i>	Text that you type	In the New Subset text box, enter <i>Proprietary Proteins.</i>
Note:	Highlights a concept of particular interest	Note: This concept is used throughout the installation manual.
Warning!	Highlights information of which you should be particularly aware.	Warning! Deleting an object will permanently delete it from the database.
{ }	Curly brackets are used for replaceable items.	Replace {root directory} with its proper value, such as c:\cabio

Table 2.1 caArray Guide Text Conventions (Continued)

CHAPTER 1 ABOUT CAARRAY

This chapter provides a caArray overview as well as a description of the relationship of caArray to caBIG®.

Topics in this chapter include:

- *caArray Overview* on this page
- *Relationship of caArray to caBIG®* on page 6

caArray Overview

caArray is an open-source, user-driven, role-based, web and programmatically accessible data management system. caArray guides the annotation and exchange of array data using a federated model of local installations whose results are sharable across the cancer Biomedical Informatics Grid (caBIG® <https://cabig.nci.nih.gov/>). Identified primarily as a data service on the Grid, caArray furthers translational cancer research through acquisition, dissemination and aggregation of semantically interoperable array data to support subsequent analysis by tools and services on and off the Grid. As array technology advances and matures, caArray will extend its logical library of assay management.

The following services are provided by caArray:

- Browsing and searching across experiments. See *Chapter 3 Navigating and Searching caArray*.
- Creating and managing array experiments. See *Chapter 4 Creating and Managing Experiments*.
- Managing array designs, protocols and vocabulary terms. See *Chapter 5 Curation Tasks*.
- Annotating experiments. See *Chapter 6 Submitting Data to an Experiment*

- Uploading, validating, and importing array data. See [Chapter 6 Submitting Data to an Experiment](#)
- Extracting data from caArray. See [Chapter 7 Extracting Data from caArray](#).
- Managing collaboration groups. See [Chapter 8 User Account Management](#).

Relationship of caArray to caBIG[®]

The National Cancer Institute (NCI) has launched the caBIG[®] (cancer Biomedical Informatics Grid[™]) initiative to accelerate research discoveries and improve patient outcomes by linking researchers, physicians, and patients throughout the cancer community.

The mission of caBIG[®] is to provide an infrastructure for creating, communicating and sharing bioinformatics tools, data and research results, using open data standards and shared data models. This supports the development of new types of analysis within and across experiments and allows new forms of collaboration, enabling the sharing of data sets and a range of analytical tools.

The primary goal of caArray is to further translational cancer research through acquisition, dissemination and aggregation of high quality array data to support subsequent analysis. The opportunity for caArray use among the cancer centers and their collaborators through caBIG[®] will ultimately benefit the cancer community.

caArray development continues to proceed with an open architecture and supportive documentation to allow for future enhancements, particularly with regard to interfacing with additional analysis tools. The goal is to create an extensible array data management system that is non-platform-specific and potentially customizable, enabling development that will continue to expand the vision of caBIG[®].

CHAPTER 2

GETTING STARTED IN CAARRAY

This chapter introduces you to the caArray interface and its navigation as well as to global operations used in all of the caArray viewing windows.

Topics in this chapter include:

- *caArray Fundamentals* on this page
- *Requesting a User Account* on page 10
- *Using caArray Online Help* on page 13
- *Navigating the caArray User Interface* on page 14

caArray Fundamentals

caArray 2.2 supports the following browsers:

- PC's running Microsoft XP: Firefox 2.0+ and Internet Explorer 6.0+
- Apple Macs running OS X: Firefox 2.0+ and Internet Explorer 5.5

If you have questions about this, contact NCICB Application Support, <http://ncicb.nci.nih.gov/NCICB/support>.

The caArray application can be accessed from NCICB using the following URL: <https://array.nci.nih.gov>.

For instructions about downloading and installing caArray 2.2 at your site, see your local administrator and/or refer to the *caArray Local Installation Guide* that can be downloaded from the caArray download page: <http://ncicb.nci.nih.gov/download/downloadcaarray.jsp>.

Note: If you are using a local installation of caArray, contact your Principal Investigator/Laboratory Manager/System Administrator for the correct URL for your use.

Browsing and Searching caArray

Browsing and Searching Before Login

Once you are at the caArray Portal Welcome login page, without being logged in you can browse caArray public data or you can perform a search of caArray public projects. (Figure 2.1).

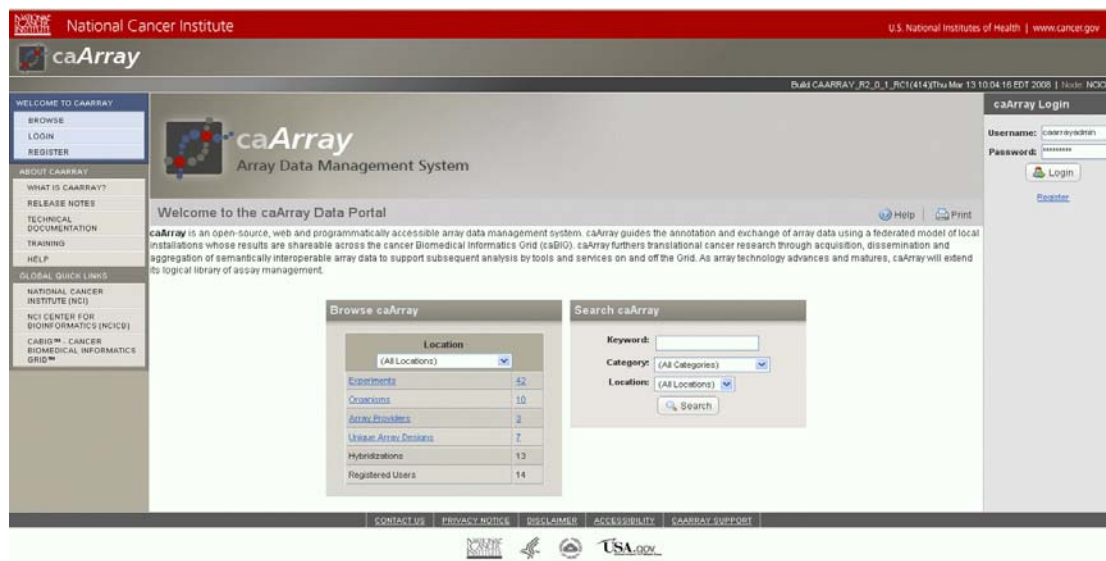


Figure 2.1 Prior to login, you can use these dialogs to browse or search the caArray database

The Browse caArray dialog box on the left center of the page lists the number of experiments saved in the database in each of the listed categories. The text categories shown in blue hypertext format can be clicked to open additional pages that display experiments with attributes in that category.

Note: The pages that open from the Browse dialog box list all caArray Experiments that have not been explicitly removed from visibility. For more information on the visibility options for Experiments, see *Experiment Visibility* on page 56.

- For more information about browsing the caArray database, see *Browsing the caArray Repository* on page 19.
- The Search caArray dialog on the right center of the page allows you to launch a search of the caArray database for public objects. For more information about executing a caArray search, see *Searching the caArray Repository* on page 22.
- Options allowing you to submit, view, modify, and add microarray experiment data to caArray are dependent on your user privileges, once you are logged in. See Table 2.2 for more information.

Note: You must obtain a user account in order to log in. For more information, see *Requesting a User Account* on page 10.

Browsing and Searching After Login

Once you have logged into caArray, the Browse and Search features are available from any page.

- To start the Browse function after login, click the **Browse** option on the left sidebar (*Figure 2.2*).

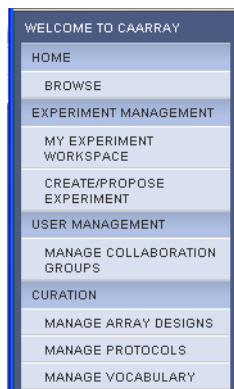


Figure 2.2 Browse options on left sidebar

caArray opens the Welcome to caArray Data Portal page, where you can launch a browse through the system, and as described in the previous section.

- To start a Search after login, enter search query parameters in the Search text box in the upper right corner of the user interface (*Figure 2.3*).



Figure 2.3 Search text box displays in every browser page

For more information about launching a search, see [Chapter 3 Navigating and Searching caArray](#).

caArray New User Account and Login

From the Welcome to caArray Data Portal page, you can register as a new user or log in if you already have a user account (*Figure 2.4*):

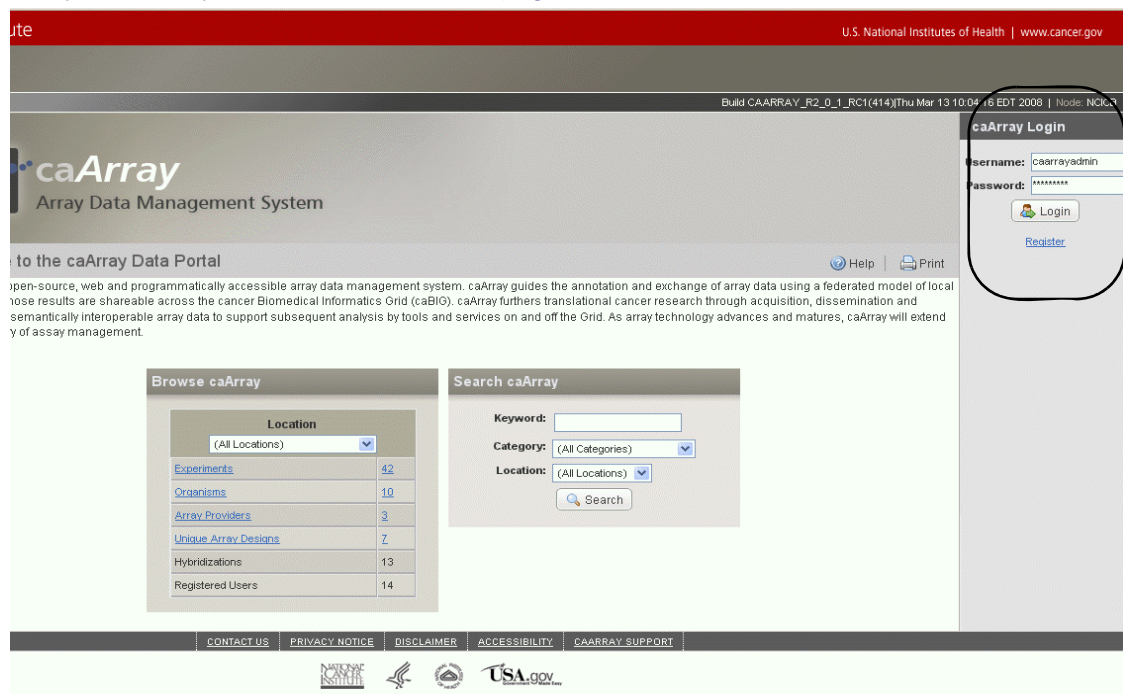


Figure 2.4 caArray login page

To register for a new user account, see the following section for more information.

To log in, follow these steps:

1. Navigate to the caArray home page. Use the URL to the NCICB instance <http://array.nci.nih.gov> (*Figure 2.4*), or contact your System Administrator for the URL to your local instance of caArray.
2. Enter your user ID and password in the upper right corner of the Welcome to caArray Data Portal page
3. Click **Login**.

After caArray verifies your credentials, the application opens to the caArray workspace, providing access to all features allowed by the permissions granted to you. For more information about finding your way around caArray, see *Navigating the caArray User Interface* on page 14.

Requesting a User Account

To request a caArray user account, you must complete steps 1 through 4.

1. Go to the NCICB caArray login page <http://array.nci.nih.gov> or use the URL provided by your System Administrator for the caArray instance at your institution.

2. Click the **Register** hypertext link, either in the left sidebar or under the caArray Login section in the upper right of the page. This opens the account registration form (*Figure 2.5*).

Figure 2.5 New user account registration form

3. In the Become a caArray User form, enter the appropriate information¹.

- **Security Information**

- **Do you have an LDAP account** [a user profile with your institution] at [NCICB or your institution]?

If **Yes**, enter your username and case-sensitive password for the purposes of verifying that it is correct. After you submit your request, you can continue to use caArray without an account to browse and search available experiments and download data while your account is verified and activated.

--Username*

--Password*

--Requested role(s)*--Select one or more of the roles. Roles are described in Table 2.1.

If your LDAP profile is not validated, caArray indicates that the LDAP credentials do not check out. You are asked to reenter them, but you can choose to answer no, and the System Administrator will manually

1. Items with an asterisk or highlight are required.

ensure you don't get a duplicate LDAP account during provisioning. You can **Cancel** or talk with your System Administrator about the problem.

If you select **No** [you do not have an LDAP account], the text boxes for entering the LDAP account information disappear. You must indicate the role you would like to be assigned in caArray, and continue entering the appropriate information in the **Account Details** section.

Role	Description	Permissible 2.0 Actions
System Administrator	Person responsible for the effective operation of caArray	Manages users
Principal Investigator [PI]	Owns experiments and studies and/or projects	Submit data Write Experiment designs Submission of annotation Submission of array data
Lab Administrator	Same as PI in caArray 2	Same as PI in caArray 2
Lab Scientist	Same as PI in caArray 2	Same as PI in caArray 2
Biostatistician	Same as PI in caArray 2	Same as PI in caArray 2
Note: In subsequent versions of caArray, role-specific features will be implemented that will limit certain actions to specific users.		

Table 2.1 caArray role descriptions

◦ **Account Details**

- **First Name***
- **Middle Initial**
- **Last Name***
- **Email [address]***
- **Organization***
- **Address [Lines 1* and 2]**
- **City***
- **State/Province***
- **Postal [or Zip] Code***
- **Country*** --Select from the drop-down list
- **Phone***
- **Fax**

4. Click **Submit Registration Request** to execute the request, or click **Cancel** to abort the registration.

Once the request for a new account is sent to NCICB, it takes 24-72 hours to process. (The process time for a local installation at your institution may differ.) You will receive an email response when the account has been activated.

Note: Once you register, you can continue to use caArray without an account to browse and search available experiments and download data while your account is activated.

When your account is registered, the UserID and password you are assigned determines your access rights for the software.

Using caArray Online Help

The caArray online help explains how to use all of the features of the caArray portal.



Note: You can open online help without being logged into caArray.

To access online help in caArray, click the **Help** icon/menu options available in the upper right corner of the user interface () and on the left sidebar under the **About caArray** section.

Online help opens with two display panels:

1. The left panel displays the Table of Contents (TOC), and also offers access to the Index and Search features of online help. The TOC can be expanded. All topics listed in the TOC and index are hypertext links to the referenced topics.
2. The right panel displays the Welcome to caArray Online Help page and other topic contents.

The following features facilitate your navigation of online help:

- The bread crumb trail at the top of the page shows the relative location of the current help topic relative to neighboring topics. Click a breadcrumb link to display that help topic.
- Click the **Back** or **Forward** links at the top of the page to display help topics you have previously viewed.
- Follow hypertext links or the **Related Topics** buttons in the help topics to open other closely related topics. If the current help page has related topics associated with it, you can also view them by clicking the **Related Topics** button () at the top right of the help page.
- Locate topics using the table of contents that displays in the left pane of the online help project or the **Index** tab that displays at the top of the Table of Contents pane.
- Perform word searches of Help by entering query text in the search text box.
- Print the current topic by clicking the **Print** button () at the top right of the help page.

Note: caArray 2.2 does not have context sensitive help.

Navigating the caArray User Interface

The caArray provides a user-friendly interface to display options relating to login, accessing documentation or performing tasks for proposed or existing experiments.

Elements in the caArray User Interface

The terms in Table 2.2 identify and describe elements of the caArray user interface.

Term	Definition
Left Vertical Navigation Task Menu	Hypertext links associated with the caArray application, caArray documentation and Global Quick Links.
[Online] Help	This caArray documentation accessed from your computer screen provides help and answers for questions about using the software. Help icon/menu options are available in the upper right corner of the user interface and on the left sidebar under the About caArray section. Note: caArray 2.2 does not have context sensitive online help. You can open online help and use the TOC, index or perform a text search.
Print	A Print icon displays on each browser interface. This prints the current page.
Browse caArray	The Browse dialog lists database categories and the number of public experiments in each. Click each hypertext link to browse details of the experiment categories.
Search {caArray database objects}	The Search caArray dialog box, available pre- and post login to all users, allows you to launch a search of public experiments. Enter keywords or select a category. (The Location dropdown lists only the current caArray instance of the users.) For more information about caArray searches, see <i>Searching the caArray Repository</i> on page 22.
Work Area Tabs	Located across the top of many of the caArray user interface windows. Work area tabs represent the tasks/annotations you create as components of an experiment
Work Queue/ Public Experiments	caArray task tabs in the My Experiment Workspace. These comprise sets of Experiment information and annotations performed on or associated with array experiments. Experiments displayed on the Work Queue tab are those with which you are associated. Experiments displayed on the Public Experiments tab are those with which you are associated and that have been made public. For more information, see <i>Experiment Visibility</i> on page 56.

Table 2.2 Components in the caArray user interface Elements of the caArray browser window

caArray Welcome Page Navigation Menu

The left sidebar of the caArray Welcome page provides links to an array of information.

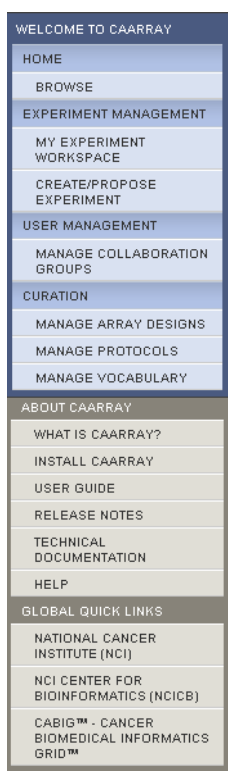


Figure 2.6 The caArray Welcome page navigation or left sidebar menu

The **Welcome to caArray** links direct you to Welcome caArray pages:

- These three options appear only before login:
 - **Login**--the Welcome/Login page
 - **Register**--the form for requesting a new account
- These four categories appear only after login:
 - **Home**--Click **Browse** to display the Browse and Search dialog boxes
 - **Experiment Management**--Open the My Experiment Workspace and the page to begin creating an Experiment.
 - **User Management**--Allows you to create and manage collaboration users groups

Note: Only System Administrators can create users. For more information, see *Managing User Accounts* on page 98.
 - **Curation**--Includes functions related to managing Array Designs and Protocols and creating and editing vocabulary terms and protocols.

The **About caArray** links direct you to various components of caArray documentation:

- **What is caArray**--Opens a brief introduction to caArray
- **Install caArray**--Opens the caArray Installation Guide pdf
- **User Guide**--opens the caArray User's Guide pdf
- **Release Notes**--opens Release Notes for caArray
- **Technical Documentation**--opens the caArray Technical Guide pdf
- **Help**--Opens the full online help project

The **Global Quick Links** provide sources for caArray-related bioinformatics information on the Internet. These include links to the following websites:

- **National Cancer Institute (NCI)** (<http://www.cancer.gov/>)
- **NCI Center for Bioinformatics (NCICB)** (<http://ncicb.nci.nih.gov/>)
- **caBIG® Cancer Biomedical Informatics Grid™** (<https://caarraydb.nci.nih.gov/caarray/>)

User Interface Footer

Options available in the footer are described as follows:

- **Contact Us**--Contact information for NCICB
- **Privacy Notice**--NIH Web Privacy Notice
- **Disclaimer**--NIH Disclaimers
- **Accessibility**--NCI Web Accessibility Feedback Form
- **User Support**--Contact information for NCICB Application Support (<http://ncicb.nci.nih.gov/NCICB/support>)

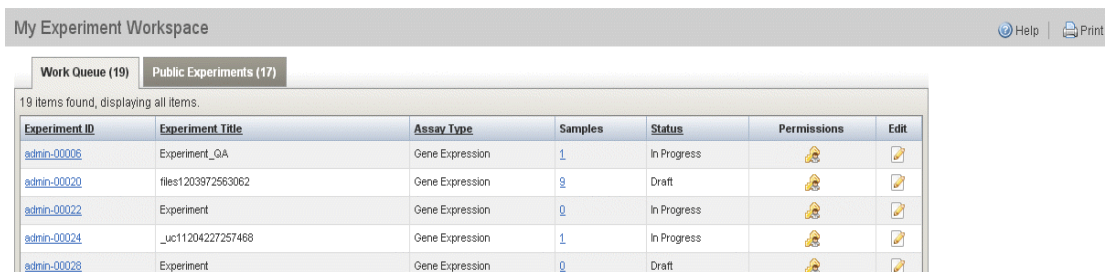
My Experiment Workspace

Note: This section describes in limited detail the elements of the experiment user interface. For more information about working with experiments, see *Chapter 3 Navigating and Searching caArray* and *Chapter 4 Creating and Managing Experiments* in this guide.

The three most common uses of caArray are described as follows:

- Submit new data, or modify existing data (with appropriate permissions)
- Search public or private data already in the caArray system
- User and group management (with appropriate permissions)

Upon login, in the user interface, the My Experiment Workspace displays tabs and options relating to draft, proposed, or public experiments with which you are associated.



The screenshot shows the 'My Experiment Workspace' interface. At the top, there are tabs for 'Work Queue (19)' and 'Public Experiments (17)'. Below the tabs, a message states '19 items found, displaying all items.' A table follows with the following data:

<u>Experiment ID</u>	<u>Experiment Title</u>	<u>Assay Type</u>	<u>Samples</u>	<u>Status</u>	Permissions	Edit
admin-00006	Experiment_OA	Gene Expression	1	In Progress		
admin-00020	files1203972563062	Gene Expression	9	Draft		
admin-00022	Experiment	Gene Expression	0	In Progress		
admin-00024	_uc11204227257468	Gene Expression	1	In Progress		
admin-00028	Experiment	Gene Expression	0	Draft		

Figure 2.7 caArray My Experiment Workspace

The My Experiment workspace displays two tabs:

1. **Work Queue**--This page lists non-public experiments with which you are associated, either in Draft or In Progress status.
2. **Public Queue**--Lists all experiments in caArray with which you are associated that have been made public.

For more information, see *Experiment Visibility* on page 56.

The listed experiments on each tab are visible in table format, according to the permissions assigned to you. The column headings display fields populated with metadata for each corresponding experiment.

Note: Column headings that are underlined are sortable by clicking on the heading.

- **Experiment ID**--The ID auto-generated by caArray; click the hypertext link to open experiment details
- **Experiment Title**--The name designated for the experiment by the principle investigator
- **Assay Type**--The category of array assay type for this experiment; either **Gene Expression** or **SNP**
- **Samples**--The number of samples used in the experiment. Click the hypertext link to open the Annotations tab to the samples details page.
- **Status**--The current status of the experiment: **Draft** or **In Progress**
- **Permissions**--Click the icon to assign or modify the experiment permissions. See *Experiment Visibility* on page 56.
- **Edit**--Click the icon to edit experiments with the appropriate permissions. See *Editing an Experiment* on page 56.

Each of these experiment elements is described in separate chapters in this user's guidelines.

Online help and additional resource links remain fixed through all the user interface pages.

CHAPTER 3

NAVIGATING AND SEARCHING CAARRAY

This chapter describes the processes for browsing and/or conducting searches of the caArray repository.

Topics in this chapter include:

- [Browsing the caArray Repository](#) on this page
-

Browsing the caArray Repository

In caArray, you can browse the repository, a feature that can be launched either before login by any user or after login by a registered user.

Note: A non-logged in user can view only public data. A logged in user can view public data and non-public data to which (s)he has assigned permissions.

You can browse experiments organized by organisms, providers, array designs from the home page of a single installation.

Once you are on the caArray Portal Welcome login page, the Browse caArray dialog box on the left center of the page lists the number of experiments saved in the database in each of the listed categories. You can click text categories shown in blue hypertext format to open additional pages that display information about those experiments ([Figure 3.1](#)).

Note: At any point in using the Browse or Search features, you can return to the home page by clicking the caArray logo at the top of the user interface or the **Browse** link in the left sidebar, if you are logged in.

From this page, you can also launch a search of the caArray repository. For more information, see .

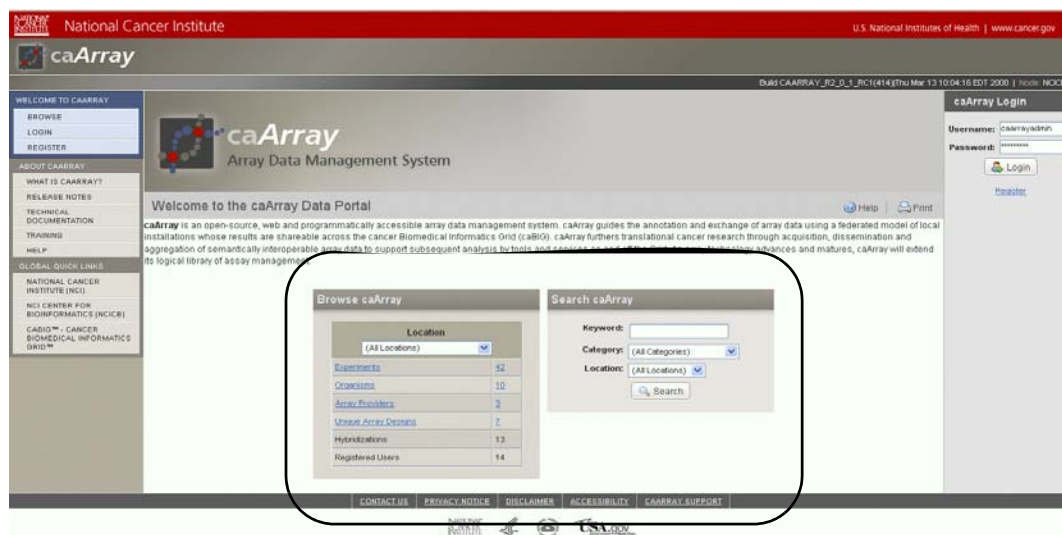


Figure 3.1 Prior to login, you can use these dialog boxes to browse or launch a search of the caArray database

To use the browse feature, follow these steps:

1. Click any of the blue hypertext links in the experiment properties category list of the Browse dialog box. The page that opens depends on the category you selected.

The descriptions in [Table 3.1](#) outline the detail that opens for each category that you click.

Browse Dialog Box Category	Description
Experiments	Both the experiments and corresponding number links open the Browse by Experiments page.
Organisms	Both the organisms and corresponding number links open the Browse by Organisms page. The page is organized alphabetically by tabs that correspond to unique organism names found among the installation's available experiments.
Array Providers	<p>Both the array providers and corresponding number links open the Browse by Array Providers page. An array provider is generally the company or group that manufactured the array design used in the experiment.</p> <p>The page is organized alphabetically by tabs that correspond to different array manufacturer's names. The count of experiments available within each named group also displays on the tab.</p> <p>Note: Only Affymetrix, Illumina and GenePix formats are fully supported with validation and parsers in caArray 2.2. For more information, see the Note about File Types in .</p>

Table 3.1 Browse dialog box categories

Browse Dialog Box Category	Description
Array Designs	An array design is a file that explains the design of an array. This includes such information as the array layout and design, its substrate, surface type, attachment type, the array strand type and the coordinates of each gene on the array.
Hybridizations	The number of hybridizations in the installation is visible, for information only. You cannot open hybridizations from this page.
Registered Users	The number of registered users in the repository is visible, for information only. You cannot open registered users from this page.

Table 3.1 Browse dialog box categories

Note: **Location** refers to the caArray instance, either at your institution or at NCICB.

2. Once the tab or page opens when you click any of these categories ([Table 3.1](#)), the same metadata displays on all pages for the list of experiments located for that category ([Table 3.2](#)).

Note: Only public experiments or non-public experiments which have not been explicitly removed from visibility display in the browse results. You can open only public experiments and non-public experiments with which you are associated..

Experiment Category	Description
<u>Experiment ID</u>	The auto-generated identification assigned by caArray. Click the hypertext link to open the corresponding experiment tabs which contain all current experiment information. Only the public data can be opened or private data to which you have been given access.
<u>Experiment Title</u>	The experiment title defined manually, naming and/or briefly describing the experiment
<u>Assay Type</u>	The type of array assay represented by the experiment; for example, Gene Expression, SNP, Exon, etc.
Primary Contact	The person named as the point of contact for the experiment. Note: The PI and POC can be the same person, but do not have to be so. Click the hypertext link or the envelope icon (✉) to open an email form where you can draft an email to this contact, if named.
<u>Organism</u>	The organism that is the source of the sample biomaterials used in the experiment
Condition/Disease State	The disease state of the source materials used in the experiment

Table 3.2 Experiment metadata categories

Experiment Category	Description
Samples	The number of samples identified in the experiment. Click the hypertext link to open the experiment to the samples details page.
<u>Updated</u>	The date of the most recent update of the experiment

Table 3.2 Experiment metadata categories

Note: Columns with underlined headings are sortable by clicking on the heading. caArray paginates the result sets in groups of 20.

Searching the caArray Repository

In caArray, you can search the repository, a feature that is available before you login or to a non-registered “Anonymous User”.

The caArray search feature allows you to locate caArray content based on user-defined search criteria. Once you find the information you seek, you can open the experiment to review or edit details, if you have proper permissions. Additionally, you can extract the data, follow hyperlinks to additional data, or you can return to the search feature to refine the query parameters.

To launch a search for a caArray experiment, follow these steps:

1. *Before you log in*, from the caArray Portal Welcome page, locate the **Search caArray** section on the lower portion of the page ([Figure 3.2](#)).

Figure 3.2 -Search dialog box

-OR-

- After you log in, locate the Search area of the page, in the upper right-hand corner.

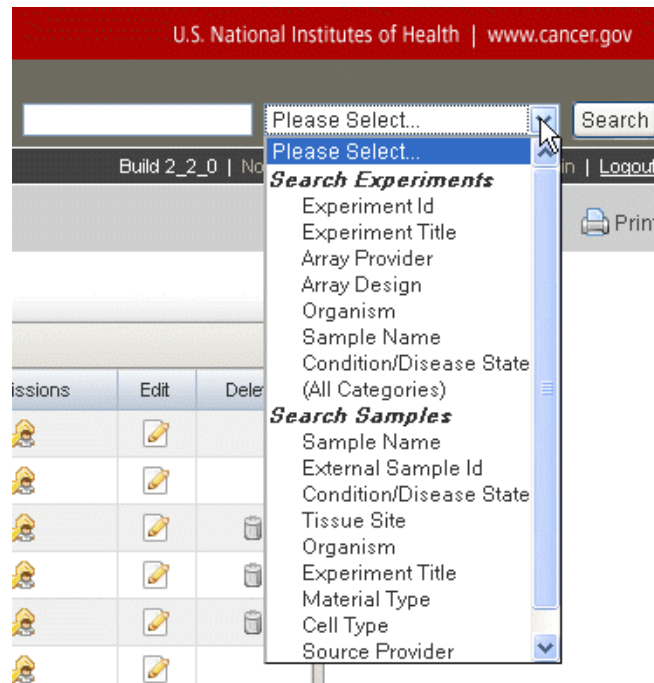


Figure 3.3 Section of every caArray page for launching a search

- Define the search criteria by using the search options described [Table 3.3](#):

Search Option	Description
Keyword	<p>In the text box, enter one or more words, separated by spaces. <i>Example: breast cancer</i></p> <p>Note: If you leave the text box empty, caArray prompts you to enter at least two characters. Queries are case insensitive; wild cards are implied on both sides of the query string. No logic statements, such as AND or OR or SQL statements are supported in these search features.</p>

Table 3.3 Search criteria options

Search Option	Description
Search Experiments – Category	<p>This option is used in conjunction with keyword(s) you enter in the Keyword field. Select one of the following Experiment property categories:</p> <ul style="list-style-type: none"> • Experiment ID • Experiment Title • Array Provider • Array Design • Organism • Sample Name • Condition/Disease State • All Categories <p>caArray searches all experiments for which the search text is present in the selected category. The search returns all experiments that match. If you do not select a category, Experiment Title remains selected (default), and the titles of all experiments is searched for the presence of the keyword.</p>
Search Samples – Category	<p>This option is used in conjunction with keyword(s) you enter in the Keyword field, except for the Organism and Other Characteristics options, which do not search with keywords. Select one of the Samples property attributes. All of the sample search options are predefined, except for Other Characteristics, described below.</p> <ul style="list-style-type: none"> • Sample Name • Sample External ID • Condition/Disease State • Tissue Site • Organism • Experiment Title • Material Type • Cell Type • Source Provider • Other Characteristics* • All Categories <p>caArray searches all samples and sources for which the search text is present in the selected category, returning all samples and/or sources that match. If you do not select a category, Sample Name remains selected (default), and all samples are searched for the presence of the keyword.</p>

Table 3.3 Search criteria options (Continued)

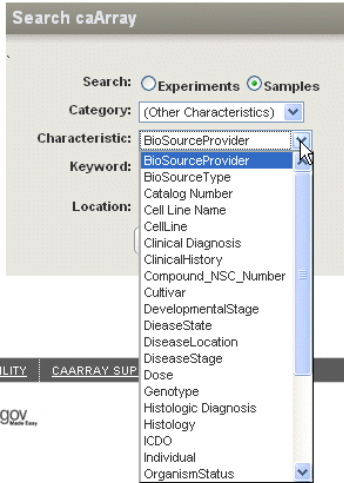
Search Option	Description
Search Samples – Category (cont'd)	<p>*When you select Other Characteristics, an additional set of attributes displays in the drop-down menu (Figure 3.4). These are arbitrary source or sample characteristics imported through MAGE-TAB SDRF files.</p> <p>Note: The Other Characteristics option is available only when you launch a search from the caArray Main Page..</p>  <p>The screenshot shows the 'Search caArray' form. The 'Search' radio buttons are set to 'Experiments' and 'Samples'. The 'Category' dropdown is set to '(Other Characteristics)'. The 'Characteristic' dropdown is open, showing a list of attributes including BioSourceProvider, BioSourceType, Catalog Number, Cell Line Name, CellLine, Clinical Diagnosis, ClinicalHistory, Compound_NSC_Number, Cultivar, DevelopmentalStage, DiseaseState, DiseaseLocation, DiseaseStage, Dose, Genotype, Histologic Diagnosis, Histology, ICDO, Individual, and OrganismStatus. The 'Keyword' and 'Location' fields are also visible.</p> <p><i>Figure 3.4 Source or Sample characteristics imported as part of MAGE-TAB SDRF file</i></p>
Location	The list displays only the current caArray instance you are using, either your local institution or CBIIT.

Table 3.3 Search criteria options (Continued)

- Click **Search** to execute the search.

Note: If you click **Search** without defining query parameters, the search is unrestricted, and all experiments in caArray that have not been explicitly removed from visibility display on the Search Results page.

Experiment Search Results

Experiment search results display on a new page, Search Results ([Figure 3.5](#)). If no results are found, a message informing you of that fact displays on the Search Results page.

Search Results							
Results for: ""							
Experiments (10)							
10 items found, displaying all items.							
Experiment ID	Experiment Title	Assay Type	Primary Contact	Organism	Condition/Disease State	# Samples	Updated
admin-00003	128CEL.zip	geneExpression	Administrator	Homo sapiens		128	1/25/2008
admin-00004	HT_HG-U133A_96F_550002	geneExpression	Administrator	Homo sapiens		0	1/25/2008
admin-00005	64 import 1201265243890	geneExpression	Administrator	Homo sapiens		0	1/25/2008
admin-00006	GSK Test	geneExpression	Administrator	Homo sapiens	Tumor Cell Line	6	1/25/2008
admin-00007	files1201267276078	geneExpression	Administrator	Homo sapiens		6	1/25/2008
admin-00008	Standard mage1201267537734	geneExpression	Administrator	Homo sapiens		26	1/25/2008
admin-00009	browsable 1201268110953	geneExpression	Administrator	Homo sapiens		0	1/25/2008
admin-00010	fractiops	geneExpression	Administrator	Hepatitis C virus		0	1/25/2008
reddy-00001	NewExperiment	geneExpression	Reddy	Gallus gallus	Squamous Carcinoma	1	1/24/2008
reddy-00002	Reddy	snp	Reddy	Gallus gallus	* Not Available	* Not Available	1/24/2008

Figure 3.5 caArray Search Results page

Note: Only public experiments or non-public experiments which have not been explicitly removed from visibility can be found via the search mechanism. You can open only public experiments and non-public experiments with which you are associated.

Search results are listed in table format, with columns displaying properties for each experiment; fields are described in [Table 3.4](#). Most of these properties were identified when the experiment was created or edited.

Note: Columns with underlined headings are sortable by clicking on the heading. caArray paginates the result sets in groups of 20.

<u>Search Results Properties</u>	Search Results Fields Descriptions
<u>Experiment ID</u>	The auto-generated identification assigned by caArray. Click the hypertext link to open the corresponding experiment tabs which contain all current experiment information.
<u>Experiment Title</u>	The experiment title defined manually, naming and/or briefly describing the experiment
<u>Assay Type</u>	The type of array assay represented by the experiment; for example, Gene Expression, SNP, Exon, etc.
<u>Primary Contact</u>	The person named as the Point of Contact for the experiment. Click the hypertext link or the envelope icon (✉) to open an email form where you can draft an email to this contact, if named.
<u>Organism</u>	The organism that is the source of the sample biomaterials used in the experiment
<u>Disease State</u>	The disease state of the source materials used in the experiment

Search Results Properties	Search Results Fields Descriptions
Samples	The number of samples identified in the experiment. If public or if you are a data owner, click the hypertext link to open the Experiment Samples tab.
<u>Updated</u>	The date of the most recent update of the Experiment draft

Table 3.4 Experiment metadata categories

You can open any experiment to which your assigned permissions grant you access. For private experiments to which you have not been assigned permission, only the ability to contact the POC is available.

To open the experiment details, click any **Experiment ID** or click the **Samples** number to open the experiment to the samples details page. You can review the experiment or contact the POC for the experiment, or with appropriate permissions, edit it or extract the experiment.

- For information about editing an experiment, see .
- For information about contacting the experiment POC, see [Primary Contact](#) in [Table 3.4](#).
- For information about extracting data from an experiment, see .

Note: At any point in using the Browse or Search features, you can return to the home page by clicking the caArray logo at the top of the page.

Sample Search Results

Sample search results display on a new page, Search Results ([Figure 3.5](#)), displaying a tab for Sample results and a tab for Source results. If no results are found, a message informing you of that fact displays on the Search Results page.

Search Results

Results for: "RNA"

Samples (1)

Sources (0)

One item found.


Sample name	External Sample Id	Description	Organism	Condition/Disease State	Tissue Site	Material Type	Cell Type	Experiment Title	Download
26-RNA		Twenty-six RNA extractions were done using standard procedures .	Homo sapiens			total_RNA		Integrated Genomics and Transcriptomic Analyses of Ductal Carcinoma In situ of the Breast	

Figure 3.6 Sample/Sources Results tab

Note: Only public samples or non-public samples which have not been explicitly removed from visibility can be found via the search mechanism. You can open only public samples and non-public samples with which you are associated. For more information about setting visibility for samples, see .

Search results are listed in table format, with columns displaying properties for each sample or source; fields are described in [Table 3.4](#). Most of these properties were identified when the sample or source was created or edited.

Note: Columns with underlined headings are sortable by clicking on the heading. caArray paginates the result sets in groups of 20.


<u>Search Results Properties</u>	<u>Search Results Fields Descriptions</u>
<u>Sample/Source Name</u>	The auto-generated identification assigned by caArray. Click the hypertext link to open the corresponding sample/source details page(s) which contain all current sample and/or source information. For more information about the sample or source tabs, see or .
Sample External ID	Identifier given to a sample in addition to its name.
Description	Description of sample or source
<u>Organism</u>	The organism that is the source of the sample biomaterials used in an experiment.
<u>Condition/Disease State</u>	The disease state of the source materials used in an experiment
<u>Tissue Site</u>	The site from which the source material was obtained.
<u>Material Type</u>	The type of source material being used for an experiment.
<u>Cell Type</u>	The category of cells used for the source material.
<u>Provider</u>	The organization that provided a source from which the sample was derived.
<u>Experiment Title</u>	The title of the experiment associated with the sample/source; defined manually, naming and/or briefly describing the experiment. Click the hypertext link to open the experiment associated with the selected sample or source.
<u>Download</u>	Click the Download icon () to download the data files associated with the sample or source. For information about extracting data from an experiment, see .

Table 3.5 Sample or Source metadata categories

Note: At any point in using the Browse or Search features, you can return to the home page by clicking the caArray logo at the top of the page.

CHAPTER 4 CREATING AND MANAGING EXPERIMENTS

This chapter describes the processes for proposing/creating caArray experiments, including all components.

Topics in this chapter include the following:

- [Overview of an Experiment](#) on this page
- [Creating an Experiment](#) on page 30
- [Updating An Experiment Proposal](#) on page 56
- [Experiment Visibility](#) on page 56

Overview of an Experiment

A caArray experiment captures all relevant information. This can include general information about the experiment, such as the experimental design and experimental factors; associated publications; biological samples; protocols; array designs; quality control and data processing steps; and so forth. Files containing the data generated for the experiment are also uploaded, validated and imported into the caArray experiment.

Once you log into caArray, the My Experiments Workspace displays by default. The basic elements of a caArray Experiment, shown on the Experiments page, are described in [Table 4.1](#):

Term	Definition
Experiment Overview	Basic information about an experiment such as IDs, service and assay types, provider of array and array designs, source of biomaterials, and disease state

Table 4.1 Elements of a caArray Experiment

Term	Definition
Contacts	Principal Investigator and/or point of contact for the experiment
Annotations	Experimental factors and design, sources, samples, extractions, labeled extracts, and hybridizations
Data	Experimental data files uploaded, validated and imported into caArray; supplemental data
Publications	Publications associated with the experiment, primarily journal articles

Table 4.1 Elements of a caArray Experiment

Managing an experiment in caArray involves two primary features:

1. Creating an experiment with appropriate characteristics and annotations. See [Creating an Experiment](#) below.
2. Uploading the experimental research data files into caArray and associating them with the appropriate samples. See [Uploading Data Files](#) on page 81.

With the appropriate permissions, you can create (“propose”) an experiment, save the draft, edit it, and finally submit an experiment with its corresponding annotations to the caArray repository. A minimum set of information must be entered for an experiment before you can save it, but almost as soon as an experiment is begun, you can save it as a draft to be retrieved and completed at a later time.

Creating an Experiment

When you create an experiment in caArray, you begin entering information on the Overview tab. Once you have saved the information on the Overview tab, the experiment becomes a draft and additional tabs for entering experiment information become available.

To create an experiment in caArray, follow these steps:

1. If you plan to use the NCICB instance of caArray, go to the NCICB caArray login page <https://array.nci.nih.gov> and log in. If you plan you use the local installation of caArray at your center, see your local System Administrator for the URL.

Once you are on the caArray Portal Welcome login page, the browser displays the experiment workspace.

2. On the left sidebar, click **Create/Propose Experiment** ([Figure 4.1](#)). This opens the Overview tab for entering overall characteristics for the experiment..



Figure 4.1 Create/Propose Experiment on left sidebar

3. Proceed to the [Overview Tab](#).

Overview Tab

When you create a new experiment in caArray, the Overview tab initially displays alone in the user interface. Once you enter the appropriate information on this tab and save it, other tabs, also necessary for adding Experiment information, display as well.

To complete the Overview tab, follow these steps:

1. On the Overview tab, enter the appropriate information for Overall Experiment Characteristics² as described in the [Table 4.2](#).

Figure 4.2 : Overview tab for an Experiment

Overview Tab Fields	Description
Experiment Title	Enter the title designated by the PI or you who are creating the experiment
Experiment Description	Enter a description for the experiment. Note: If you import MAGE-TAB data into your experiment, the description you enter here will be overwritten by the one in the MAGE-TAB IDF.
Status	Draft displays by default before the experiment is formally submitted.

Table 4.2 Fields for Overall Experiment Characteristics

2. Fields with a red asterisk * are required.

Overview Tab Fields	Description
Experiment Identifier	<p>This project identifier is autogenerated by caArray upon the initial save of the experiment. The experiment identifier is not editable. The ID is generated using the PIs last name followed by a 5 character number. <i>Example:</i> jdoe-90765. After the experiment has been saved or submitted, the experiment ID displays as a hypertext link that opens the experiment.</p> <p>Note: When the experiment is a <i>Draft</i>, if the PI's name is changed under Contacts (see <i>Contacts Tab</i> on page 33), this Experiment ID changes accordingly. However, once the experiment is <i>In Process</i>, even if the PI for the experiment is changed, the Experiment ID does not change. This is to ensure a stable accession number for the experiment.</p>
Assay Type	<p>Select from the drop-down menu the appropriate assay type. Options are the following:</p> <ul style="list-style-type: none"> • Gene Expression--experiment using microarrays intended to measure levels of transcribed genes • SNP--experiment using microarrays intended to detect nucleotide changes in chromosomal DNA • aCGH--array comparative genomic hybridization; a method for the analysis of chromosome copy number changes (gains/losses). • Exon--Exon arrays are designed to study which exons are present in an expressed gene. • microRNA--Experiment that measures activity among the genes encoding miRNA. • Methylation--experiment that attempts to establish patterns of methylation genome-wide or within targeted promoters or CpG islands
Provider	<p>Select from the drop-down menu the provider of the array.</p> <p>Note: Only Affymetrix, Illumina and GenePix formats are fully supported with validation and parsers in caArray 2.2. For other providers, data files are stored in the database in their native format only. For more information, see the Note about File Types in <i>Managing Data</i> on page 77.</p> <p>Once selected, caArray automatically loads a corresponding list of array designs (next field).</p>
Array Designs	<p>Select one or multiple array designs, (using CTRL + click or SHIFT + click), from the automatically-generated list of array designs corresponding to the provider you selected. The array design of interest may have already been imported into caArray, or you can choose to import those of your choice. For more information, see <i>Managing Array Designs</i> on page 64.</p>
Organism	<p>Select from the drop-down menu the organism that is the source of the sample biomaterial used in the experiment.</p>

Table 4.2 Fields for Overall Experiment Characteristics

2. After entering the information, click the **Save** button at the bottom of the page. Upon saving, caArray validates required fields and saves the experiment as a draft. A confirmation messages displays, verifying that the proposal is saved. If the validation fails, caArray display a message indicating which field(s) need correction.

When you save the draft successfully, other tabs used for adding additional information for the experiment display.

3. Proceed to the **Contacts** tab [Contacts Tab](#).

Contacts Tab

A caArray contact can be a principal investigator (PI), the point of contact (POC), or in any other way associated with the experiment, such as a biomaterial provider, consultant, etc. The contact does **not** have to be a registered user of caArray.

By default, the person who creates an experiment is listed on the Contacts tab ([Figure 4.3](#)).

To enter contact information for the experiment, on the Contacts tab follow these steps:

1. Click the **Add a New Contact** button.
1. Enter information for the fields described in [Table 4.3](#).



Figure 4.3 A cropped version of the Contacts tab

Contact Fields	Description
First Name*	First and last names of the contact. Note that the contact does not have to be a registered user of caArray.
Last Name*	
Email*	Email address of the contact
Phone	Phone number of the contact

Table 4.3 Contact fields

Contact Fields	Description
Roles*	<p>Select the role performed by the contact you are adding and click the adjoining icon (+) to move it into the Selected Roles panel.</p> <p>If the appropriate value is not displayed, to find a role of interest that might already be in caArray, begin typing a term in the Filter text box. The available roles that display in the panel below are limited according to the text you enter. A message displays if a corresponding role cannot be found.</p>

Table 4.3 Contact fields

- Click **Save**.
- A contact can be edited or deleted. To do so, click the **Edit** or **Delete** button corresponding to the contact on the Contacts tab. Proceed with the edit as described above for creating the contact.
- Proceed to the [Annotations Tab](#)Annotations tab.

Annotations Tab

The Annotations tab opens with seven subtabs for entering annotation data for the experiment you are creating. The browser displays the Experimental Design subtab by

default ([Figure 4.4](#)). Enter the appropriate information as described below in the following topics for each of the seven subtabs.

Experiment: H_JB123KLH4

Experiment: H_JB123KLH4

Overview | Contacts | **Annotations** | Data | Parameters

Experimental Design | Experimental Factors | Sources | Samples | Extracts | Labeled Extracts | Hybridizations

Experimental Design

Required fields are marked with ***asterisks***.

Experiment Design Types:

Filter:

- ☒ all_pairs (MO)
- ☒ array_platform_variation_design (MO)
- ☒ binding_site_identification_design (MO)
- ☒ cellular_modification_design (MO)
- ☒ cellular_process_design (MO)

Selected Experiment Design Types

Experiment Design Description:

Quality Control Types:

Filter:

- ☒ biological_replicate (MO)
- ☒ dye_swap_quality_control (MO)
- ☒ peer_review_quality_control (MO)
- ☒ real_time_PCR_quality_control (MO)
- ☒ reverse_transcription_PCR_quality_control

Selected Quality Control Types

Quality Control Description:

Replicate Types:

Filter:

- ☒ biological_replicate (MO)
- ☒ dye_swap_replicate (MO)
- ☒ technical_replicate (MO)

Selected Replicate Types

Replicate Description:

Figure 4.4 Upon opening, the Annotations tab displays the Experimental Design subtab and 6 other subtabs for entering experiment annotation data.

Experimental Design

The experimental design, ([Figure 4.4](#)), describes the intent of the research and a description that is common to all hybridizations performed in the experiment.

1. Enter on the Experimental Design subtab (under the Annotations tab) the appropriate information as described in [Table 4.4](#)³.

Experimental Design Fields	Description
Experimental Design Type*	<p>If the appropriate Experimental Design Type displays in the left panel, click the adjoining icon (+) to move it into the Selected Experimental Design Types panel.</p> <p>If the appropriate value is not displayed, to find a design type of interest that might already be in caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p>
Experimental Design Description	Enter a description for the experimental design used for the experiment.
Quality Control Types	<p>Select the QC type in the displayed list.</p> <p>If the appropriate QC Type displays in the left panel, click the adjoining icon (+) to move it into the Selected QC Types panel.</p> <p>If the appropriate value is not displayed, to find a QC type of interest that might already be in caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p>
Quality Control Description	Enter a description for the quality control used for the experiment.
Replicate Types	<p>Select one or more replicate types from the displayed list. Replicates can be either technical (arrays) or biological (laboratory animals or samples, etc.)</p> <p>If the appropriate Replicate Type displays in the left panel, click the adjoining icon (+) to move it into the Selected Replicate Types panel.</p> <p>If the appropriate value is not displayed, to find a replicate type of interest that might already be in caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p>
Replicate Description	If there are replicates used in the experiment, describe the number of replications and how the replicates were generated.

Table 4.4 Experimental Design fields

3. Fields with a red asterisk * are required.

Note: **Experiment Design Type**, **Quality Control Type**, and **Replicate Type** terms are included from the MGED Ontology, <http://mged.sourceforge.net/ontologies/MGEDontology.php>, by default, but at the present time it is possible to add additional (non-MO) terms via MAGE-TAB import. Therefore, you may see non-MO terms in these lists. See *Importing MAGE-TAB Data* on page 89.

2. Click **Save** to save the draft. Click **Cancel** to return to the subtab without adding the design.
3. Proceed to the *Experimental Factors* **Experimental Factors** subtab.

Experimental Factors

Experimental factors are the intended sources of variation in the experiment. The Experimental Factors subtab (under the Annotations tab) displays any previous Experimental factors that have been added to the experiment you are creating.

1. On the Experimental Factors subtab, click the **Add a New Experimental Factor** button.
2. In the form that opens, enter the information as described in *Table 4.5*⁴.

<i>Experimental Factors Fields</i>	<i>Description</i>
Factor Name*	Enter a name for the experimental factor.
Description	Enter a description for the experimental factor.
Category	Select the appropriate category for the experimental factor in the displayed list. Note: Terms are included from the MGED Ontology, http://mged.sourceforge.net/ontologies/MGEDontology.php by default, but at the present time it is possible to add additional (non-MO) terms via MAGE-TAB import. Therefore, you may see non-MO terms in these lists.

Table 4.5 Experimental Factor fields

3. Click **Save** to save the draft. Click **Cancel** to return to the subtab without adding the factor.
4. Repeat steps 1 - 3 as often as needed to enter all the experimental factors for this experiment.
5. Click **Save** or **Cancel** to abort the action.
6. Proceed to the Sources subtab. See also *Biological Source Material* on this page.

Biological Source Material

Materials of biological origin are used in array experiments, and the state and characteristics of those biomaterials can be objectively documented and described. Biological materials can be treated, extracted, labeled and hybridized on arrays to

4. Fields with a red asterisk * are required.

study the characteristics of the nucleic acid sequences represented on the arrays. All categories of biological materials are created in caArray to be included as essential components in a caArray experiment.

Many biological materials' characteristics are defined in caArray by terms found in the caArray Controlled Vocabulary Terms. For more information, see *Managing [Controlled] Vocabulary [Terms]* on page 72.

In caArray, biological materials are divided into four different categories based on the treatment status of the material. The categories are based on MAGE-TAB specifications, as described in this paper: <http://www.biomedcentral.com/1471-2105/7/489>. The biological materials consist of **Sources**, **Samples**, **Extracts** and **Labeled Extracts**, defined below and illustrated in *Figure 4.5* in their hierarchical relationship.

- **Source** is any biological site from which the tissue for the array is derived before any preparation of the tissue for the array takes place. *Example:* human brain tumor tissue that has been treated with an anti-cancer drug.
- **Samples** are the original source biomaterials after initial treatment events. *Example:* A tissue or biopsy material treated to create a cell lysate.
- **Extracts** are samples after a treatment event in which DNA or RNA is extracted. *Example:* A sample tissue treated with RNA extraction method yields the extract, RNA.
- **Labeled Extracts** are extracts that have been labeled for detection of the nucleic acids on the array. *Example:* The extract RNA is labeled with a fluorescent dye, yielding the labeled extract RNA.

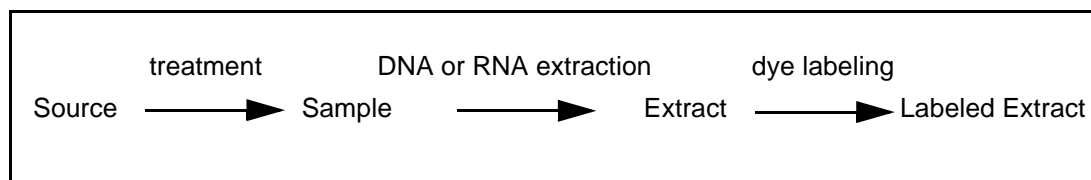


Figure 4.5 Biomaterials components and their relationship in caArray

The nature of each type of biomaterial is defined in varying dialog boxes in caArray.

Biomaterials can be created manually as described in the following sections . Alternatively, they can be generated automatically when data files are imported into caArray. For more information, see *Importing Data* on page 86. When data files are thus associated with biomaterials, the files can be downloaded from each of the corresponding biomaterial tabs. For more information, see the [biomaterials] tab sections and *Downloading Data from caArray* on page 93.

Sources Tab

A source is any biological site from which the tissue for the array is derived before any preparation of the tissue for the array takes place. *Example:* human brain tumor tissue that has been treated with an anti-cancer drug.

The Sources subtab (under the Annotations tab) displays any sources that have been added to the experiment you are creating.

To add a Source, follow these steps:

1. On the Sources tab, click the **Add a new Source** button in the upper right-hand corner of the tab. This opens the Sources page for adding source information [Figure 4.6](#)).

[Sources](#) > Add a new Source

Required fields are marked with **'asterisks'**.

Source name*:

Description:

Tissue Site*:

- ☐ A qa tissue site 2.1 alpha 1 (ALPHA3_QA)
- ☐ ALPHA4 (NCBI Taxonomy)
- ☐ AR_ALPHA3 (caArray)
- ☐ Bladder (NCI_Thesaurus)
- ☐ Bladder (NCI Thesaurus)

Selected Tissue Site

Material Type:

- ☐ A qa material type 2.1 alpha 1 (qa source)
- ☐ ALPHA4_ARTI (The Unified Code for Units of Measure)
- ☐ alpha4_value (caArray)
- ☐ ALPHA_Material_Types (MO)

Selected Material Type

Cell Type:

- ☐ A qa cell type 2.1 alpha 1 (CTO)
- ☐ ALPHA4 (ArrayExpress)
- ☐ astrocyte (NCI_Thesaurus)
- ☐ Breast (caArray)
- ☐ B_lymphoblast (CTO)

Selected Cell Type

Disease State:

- ☐ A qa condition/disease state 2.1 alpha 1 (CTO)
- ☐ Acute B Cell Lymphocytic Leukemia (NCI Metathesaurus)
- ☐ Acute B Cell Lymphocytic Leukemia (caArray)

Selected Disease State

Protocol Type: --Select a Protocol Type--

Protocols:

-- No items found --

Selected Protocols

Drag items to reorder list

Figure 4.6 Sources subtab

2. In the Sources form, enter the information as described in [Table 4.6](#).⁵

Source Fields	Description
Source Name*	Name assigned to the source
Description	Description of the source

Table 4.6 Fields for documenting a source

5. Fields with a red asterisk * are required.

Source Fields	Description
Tissue Site*	Tissue site is the site from which the source material was obtained. You can choose from available terms or add a new term. For more information about adding a new term to annotate this attribute, see <i>Adding Vocabulary for Experiments</i> on page 51.
Material Type	Material type is the descriptor for the type of source material being used for the experiment. You can choose from available terms or add one or more new terms. caArray comes pre-loaded with MO terms for Material Type - this is the preferred vocabulary. There are three ways you can enter terms for annotating this attribute. See <i>Adding Vocabulary for Experiments</i> on page 51 for more information about using this feature.
Cell Type	Cell type is the descriptor for the category of cells used for the source material. You can choose from available terms or add one or more new terms. For more information about adding a new term to annotate this attribute, see <i>Adding Vocabulary for Experiments</i> on page 51.
Disease State	Disease state is the descriptor used to describe the disease condition when the source was obtained. You can choose from available terms or add one or more new terms. For more information about adding a new term to annotate this attribute, see <i>Adding Vocabulary for Experiments</i> on page 51.

Table 4.6 Fields for documenting a source

- Click **Save** to save the source to the experiment. Click **Cancel** to return to the subtab without adding the source.

Note: When you create samples for this experiment, you must specify the source(s) for the samples. The Samples column on this Sources tab will then be populated.

- Repeat steps 1-3 as often as necessary to add all appropriate sources to the experiment.
- Proceed to the **Samples** Tab.

See *Copying a Biomaterial/Hybridization* on page 50, *Deleting a Biomaterial/Hybridization* on page 50, and *Downloading Associated Data Files* on page 50. Biomaterials and hybridizations can also be edited; for more information, see *Editing Experiment Annotations* on page 49.

Samples Tab

A sample is the original source biomaterial after initial treatment events. *Example:* A tissue or biopsy material treated to create a cell lysate. The Samples subtab (under the Annotations Tab) displays any previous samples that may have been added previously to the experiment you are creating.

The Samples subtab (under the Annotations Tab) displays samples that have been added to the experiment you are creating.

1. On the Samples tab, click the **Add a New Sample** button to add a new sample. This opens the Samples page when you can add sample information (Figure 4.7). The Extracts and Labeled Extracts pages are similar to this page.

[Samples](#) > Add a new Sample

Required fields are marked with ***asterisks***.

Sample name*:

Description:

External ID:

Sources*: Filter:

- 10219-10218 #1656 GBM 133A 8-13-02
- joh test annot
- source joh 1

Selected Sources

Material Type: Filter:

- A qa material type 2.1 alpha 1 (qa source)
- ALPHA4_ARTI (The Unified Code for Units of Measure)
- alpha4_value (caArray)
- ALPHA_Material_Types (MO)

Selected Material Type

Protocol Type: --Select a Protocol Type--

Protocols: Filter:

-- No items found --

Selected Protocols

Figure 4.7 A portion of a Samples page

2. In the Sample form, enter the information described in Table 4.7.⁶

Samples Fields	Description
Sample Name*	Enter a name for the sample.
Description	Enter a description of the sample.
External ID	Enter an additional identifier given to the sample, beyond the sample name. You can create this ID here or you can add it as a field in a MAGE-TAB SDRF you import into this experiment. If you try to enter the same external Sample ID for two different samples within the same Experiment, caArray disallows it.

Table 4.7 Fields for documenting samples

6. Fields with a red asterisk * are required.

Samples Fields	Description
Source(s)*	<p>Sources must already have been saved to caArray. Select one or more sources from which the sample was derived. As you do so, the selected source(s) move into the Selected Sources panel.</p> <p>If the appropriate value is not displayed, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found. See <i>Sources Tab</i> on page 38 for more information.</p>
Material Type	<p>Material type is the descriptor for the type of source material being used for the experiment.</p> <p>You can choose from available terms or add one or more new terms. caArray comes pre-loaded with MO terms for Material Type - this is the preferred vocabulary.</p> <p>There are three ways you can enter terms for annotating this attribute. See <i>Adding Vocabulary for Experiments</i> on page 51 for more information about using this feature.</p>
Protocol Type	<p>Protocol Type terms listed are MGED Ontology terms that come pre-loaded with caArray, but additional terms can be added on the Manage Protocols page (<i>Managing Protocols</i> on page 68) or via MAGE-TAB (<i>Importing MAGE-TAB Data</i> on page 89).</p>
Protocol	<p>If the appropriate protocol displays in the list, click the adjoining icon (⊕) to move it into the Selected Protocols panel. Note: The available selections are limited based on the protocol type selected above.</p> <p>If the appropriate value is not displayed, to find a protocol of interest that might already be in the caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p> <p>If the appropriate protocol has not been entered into the system, click Add to open the page where you can add a new protocol. For more information, see <i>Creating a Protocol</i> on page 70.</p>

Table 4.7 Fields for documenting samples

- Click **Save**. Click **Cancel** to return to the subtab without adding the sample.

Note: When you create extracts for this experiment, you must specify the samples for the extracts. The Extracts column on this Samples tab will then be populated.

- Repeat steps 1 - 3 as often as needed to enter all the samples used in this experiment.
- Proceed to the **Extracts** tab.

See *Copying a Biomaterial/Hybridization* on page 50, *Deleting a Biomaterial/Hybridization* on page 50, and *Downloading Associated Data Files* on page 50. Biomaterials and hybridizations can also be edited; for more information, see *Editing Experiment Annotations* on page 49.

Extracts Tab

An extract is a sample after a treatment event in which DNA or RNA is extracted for the array.

The Extracts subtab (under the Annotations tab) displays extracts that have been added to the experiment you are creating.

1. On the Extracts subtab, click the **Add a New Extract** button to add a new extract. This opens the Labeled Extracts page that is similar to the Samples page (*Figure 4.7*).
2. In the Extract form, enter the information described in *Table 4.8*.⁷

<i>Extracts Fields</i>	<i>Description</i>
Extract Name*	Name assigned to the extract
Description	Description of the extract
Samples*	<p>Samples must already have been saved to caArray. Select one or more samples from which the extract was derived. As you do so, the selected sample(s) move into the Selected Samples panel.</p> <p>If the appropriate value is not displayed, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found. See <i>Samples Tab</i> on page 40 for more information.</p>
Material Type	<p>Material type is the descriptor for the type of source material being used for the experiment.</p> <p>You can choose from available terms or add one or more new terms. caArray comes pre-loaded with MO terms for Material Type - this is the preferred vocabulary.</p> <p>There are three ways you can enter terms for annotating this attribute. See <i>Adding Vocabulary for Experiments</i> on page 51 for more information about using this feature.</p>
Protocol Type	Protocol Type terms listed are MGED Ontology terms that come pre-loaded with caArray, but additional terms can be added on the Manage Protocols page (<i>Managing Protocols</i> on page 68) or via MAGE-TAB (<i>Importing MAGE-TAB Data</i> on page 89).

Table 4.8 Fields for documenting an extract

7. Fields with a red asterisk * are required.

Extracts Fields	Description
Protocol	<p>If the appropriate protocol displays in the list, click the adjoining icon (+) to move it into the Selected Protocols panel. Note: The available selections are limited based on the protocol type selected above.</p> <p>If the appropriate value is not displayed, to find a protocol of interest that might already be in the caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p> <p>If the appropriate protocol has not been entered into the system, click Add to open the page where you can add a new protocol. For more information, see <i>Creating a Protocol</i> on page 70.</p>

Table 4.8 Fields for documenting an extract

- Click **Save**. Click **Cancel** to return to the subtab without adding the Extract.
- Note:** When you create labeled extracts for this experiment, you must specify the extracts for the labeled extracts. The labeled extracts column on this Extracts tab will then be populated.
- Repeat steps 1 - 3 as often as needed to enter all the extracts used in this experiment.
 - Proceed to the Labeled Extracts tab.

See *Copying a Biomaterial/Hybridization* on page 50, *Deleting a Biomaterial/Hybridization* on page 50, and *Downloading Associated Data Files* on page 50. Biomaterials and hybridizations can also be edited; for more information, see *Editing Experiment Annotations* on page 49.

Labeled Extracts Tab

A labeled extract is an extract that has been labeled for detection of the nucleic acids on the array. *Example:* The extract RNA is labeled with a fluorescent dye, yielding the labeled extract RNA.

The Labeled Extracts subtab (under the Annotations tab) displays labeled extracts that have been added to the experiment you are creating.

- Click the **Add a New Labeled Extract** button to add a new labeled extract. This opens the Labeled Extracts page that is similar to the Samples page (Figure 4.7).
- In the Labeled Extract form that opens, enter the information described in Table 4.9⁸.

Labeled Extracts Fields	Description
Labeled Extract Name*	Name assigned to the extract

Table 4.9 Fields for documenting a labeled extract

Labeled Extracts Fields	Description
Description	Description of the extract
Extracts*	<p>Extract(s) from which the labeled extract was derived. Extracts must already have been saved to caArray.</p> <p>Select one or more extracts from which the labeled extract was derived. As you do so, the selected extract(s) move into the Selected Extracts panel.</p> <p>If the appropriate value is not displayed, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p> <p>See <i>Extracts Tab</i> on page 43 for more information.</p>
Material Type	<p>Material type is the descriptor for the type of source material being used for the experiment.</p> <p>You can choose from available terms or add one or more new terms. caArray comes pre-loaded with MO terms for Material Type - this is the preferred vocabulary.</p> <p>There are three ways you can enter terms for annotating this attribute. See <i>Adding Vocabulary for Experiments</i> on page 51 for more information about using this feature.</p>
Protocol Type	<p>Protocol Type terms listed are MGED Ontology terms that come pre-loaded with caArray, but additional terms can be added on the Manage Protocols page (<i>Managing Protocols</i> on page 68) or via MAGE-TAB (<i>Importing MAGE-TAB Data</i> on page 89).</p>
Protocol	<p>If the appropriate protocol displays in the list, click the adjoining icon (⊕) to move it into the Selected Protocols panel. Note: The available selections are limited based on the protocol type selected above.</p> <p>If the appropriate value is not displayed, to find a protocol of interest that might already be in the caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p> <p>If the appropriate protocol has not been entered into the system, click Add to open the page where you can add a new protocol. For more information, see <i>Creating a Protocol</i> on page 70.</p>

Table 4.9 Fields for documenting a labeled extract

- Click **Save**. Click **Cancel** to return to the subtab without adding the labeled extract.

Note: When you create hybridizations for this experiment, you must specify the labeled extracts for the Hybridizations. The Hybridizations column on this Labeled Extracts tab will then be populated.

- Fields with a red asterisk * are required.

4. Repeat steps 1 - 3 as often as needed to enter all the labeled extracts used in this experiment.
5. Proceed to the *Hybridizations Tab*.

Hybridizations Tab

In caArray, a hybridization is an array with which one or more labeled extracts has been incubated. Using this technique, single stranded nucleic acids are allowed to interact so that complexes, or hybrids, are formed by molecules with sufficiently similar, complementary sequences. By this means the degree of sequence identity can be assessed and specific sequences detected.

The Hybridizations subtab (under the Annotations tab) displays hybridization information that has been added to the experiment you are creating.

1. Click the **Add a New Hybridization** button to add a new hybridization. This opens the Hybridizations page where you can add hybridization information (*Figure 4.8*).

Figure 4.8 Hybridizations page

2. In the Hybridizations form, enter the information described in *Table 4.10*⁹.

<i>Hybridizations Fields</i>	<i>Description</i>
Hybridization Name*	Name assigned to the hybridization
Description	Description of the hybridization

Table 4.10 Fields for documenting a hybridization

⁹ Fields with a red asterisk * are required.

Hybridizations Fields	Description
Labeled Extracts*	<p>Labeled extract(s) used in the hybridization protocol. Select one or more listed labeled extracts. As you do so, the selected labeled extract(s) move into the Selected Labeled Extracts panel.</p> <p>If the appropriate value is not displayed, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found. See <i>Labeled Extracts Tab</i> on page 44 for more information.</p>
Array Designs	<p>This field displays only if you associated more than one array design on the <i>Overview Tab</i>, described on page 31. Select the array design appropriate for this hybridization.</p>
Protocol Type	<p>Protocol Type terms listed are MGED Ontology terms that come pre-loaded with caArray, but additional terms can be added on the Manage Protocols page (<i>Managing Protocols</i> on page 68) or via MAGE-TAB (<i>Importing MAGE-TAB Data</i> on page 89).</p>
Protocol	<p>If the appropriate protocol displays in the list, click the adjoining icon (⊕) to move it into the Selected Protocols panel. Note: The available selections are limited based on the protocol type selected above.</p> <p>If the appropriate value is not displayed, to find a protocol of interest that might already be in the caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p> <p>If the appropriate protocol has not been entered into the system, click Add to open the page where you can add a new protocol. For more information, see <i>Creating a Protocol</i> on page 70.</p>

Table 4.10 Fields for documenting a hybridization

- Click **Save**. The new hybridization object displays on the Hybridization tab. Click **Cancel** to return to the subtab without adding the hybridization.
- Note:** On the saved draft, the uncompressed size of the hybridization file is defined by caArray when it becomes available.
- Repeat steps 1 - 3 as often as needed to enter all the hybridization data used in this Experiment.
 - Proceed to the *Data Tab* on page 53.

Viewing a Hybridization

1. To open and view the information for a hybridization, click on a selected hybridization hypertext link in the **Hybridization Name** column on the the Hybridization tab (circled in [Figure 4.9](#)).

Experiment: A catalytic antioxidant (AEOL 10150) attenuates expression of inflammatory genes in stroke.

Overview

Contacts

Annotations

Data

Publications

Experimental Design

Experimental Factors

Sources

Samples

Extracts

Labeled Extracts

Hybridizations

Hybridizations

17 items found, displaying 1 to 15. [FirstPrev] 1, 2[NextLast]

Hybridization name	Related Labeled Extract	Uncompressed Size	Edit
gov.nih.nci.ncicb.caarray:H...	Biotin labeled cRNA fro B02...	237	
gov.nih.nci.ncicb.caarray:H...	Biotin labeled cRNA fro B02...	237	
gov.nih.nci.ncicb.caarray:H...	Biotin labeled cRNA fro B02...	239	
gov.nih.nci.ncicb.caarray:H...	Biotin labeled cRNA fro B02...	238	
gov.nih.nci.ncicb.caarray:H...	Biotin labeled cRNA fro B02...	238	

Figure 4.9 Hybridizations associated with an experiment appear on the Hybridizations tab

The page that opens displays hybridization details. This includes experimental factors and their values and any data files that have been uploaded and associated with the experiment ([Figure 4.10](#)).

Experiment Details Submit Experiment F

Experiment: A catalytic antioxidant (AEOL 10150) attenuates expression of inflammatory genes in stroke.

Overview	Contacts	Annotations	Data	Publications		
Experimental Design	Experimental Factors	Sources	Samples	Extracts	Labeled Extracts	Hybridizations

[Hybridizations](#) > gov.nih.nci.ncicb.caarray:Hybridization:1015897590474569:1

Required fields are marked with ***asterisks***.

Description:
Labeled Extracts*: Biotin labeled cRNA fro B029_Brain
Array Design: mg_u74av2
Protocols: Affymetrix MG_U74Av2 Feature Extraction Suite 4.0

Edit

Values Experimental factor name(s) and value(s)

Factor Name	Factor Value
Treatment	sham surgery

Download Data Data files associated with the experiment

Filter By File Type:	(All)	Filter By File Type:	(All)
	File Name	File Type	Ext.
	GSM2206.bt	AFFYMETRIX_TXT	.bt
		Compressed Size	Uncompressed Size
		74 KB	239 KB

Download Queue [Show Files](#)
 Job Size: 0 Files, 0 KB

Clear Download Queue Launch Download Job

Figure 4.10 Experiment Details, Hybridization page

2. To open experimental factor details, click on the link in that section of the page. A page showing all experimental factors associated with the experiment opens.

3. You can filter the data files by file type or file status. To apply the filter(s), click the drop-down arrows to select the filter criteria. As soon as you make the selection, the list is filtered.
4. Click the **Download Job** button to download all files that are in the Download Data list.
5. When the download finishes, specify in the dialog box opens to open or to save the file.



Managing Annotations

Once annotations have been added to an experiment, you can perform several annotation-related tasks. See the following topics for more information.

- *Editing Experiment Annotations* on page 49
- *Copying a Biomaterial/Hybridization* on page 50
- *Deleting a Biomaterial/Hybridization* on page 50
- *Downloading Associated Data Files* on page 50
- *Adding Vocabulary for Experiments* on page 51


Editing Experiment Annotations

Once an experiment is in draft or in progress status (having been submitted), as the experiment creator or having Write access as a collaborator, you can add data such as annotations, files and publications to the experiment. To do so, follow these steps:

1. In the My Experiment Workspace, locate the experiment of interest on the **Work Queue** tab.
2. Click the **Edit** button () that corresponds to the selected experiment.
3. To change an attribute, navigate to the appropriate tab. In some instances, such as on the Experimental Design tab, you can directly remove attributes by clicking the () button. In other instances, such as on a biomaterials tab, you can click an **Edit** button that takes you to another page where you can edit the item as just described. Or you can click a **Delete** icon corresponding to an item you want to delete.
4. To add an attribute, navigate to the tab for the attribute or item you want to add. The tabs where you can add data have an **Add {attribute}** button in the upper right corner. To mention a few, you could add samples, labeled extracts, upload data or associate publications with your experiment.
5. Enter the appropriate information you wish to add. For more specific information, see *Creating an Experiment* on page 30. See also *Editing an Experiment* on page 56.
6. Click **Save** to complete adding the data.

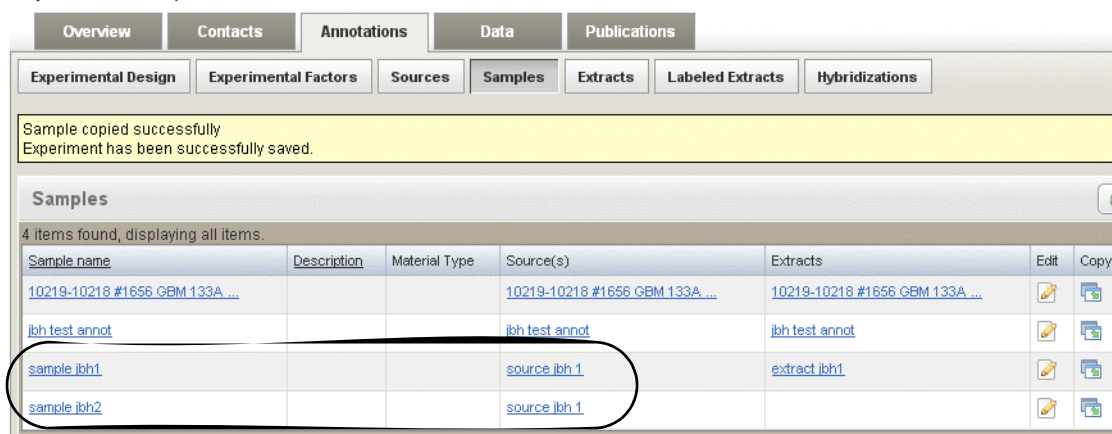
For details about each tab where you can edit annotations, see topics under *Annotations Tab* on page 34.

Copying a Biomaterial/Hybridization

To copy a source, sample, extract, labeled extract or hybridization, click the **Copy** icon () that corresponds to the biological source material or hybridization on its tab under Annotations in the experiment.

caArray copies the biomaterial/hybridization attributes, renames the copied object using the existing source name and adding an incremental number. The copied source now displays under the original (*Figure 4.11*).

Experiment: test jbh












Sample name	Description	Material Type	Source(s)	Extracts	Edit	Copy
10219-10218 #1656 GBM 133A ...			10219-10218 #1656 GBM 133A ...	10219-10218 #1656 GBM 133A ...		
jbh test annot			jbh test annot	jbh test annot		
sample jbh1			source jbh 1	extract jbh1		
sample jbh2			source jbh 1			

Figure 4.11 A biomaterial or hybridization can be copied in an experiment. The last sample in this list was copied from the previous sample. The name for the sample is automatically generated by caArray. Note that the copied sample originates from the same source as sample jbh1.

Deleting a Biomaterial/Hybridization

Biomaterials and hybridizations can be deleted, but the biomaterial chain should be considered. If a source is associated to a sample, the sample must be deleted first and so forth down the chain.

To delete a source, sample, extract, labeled extract or hybridization, click the **Delete** icon () that corresponds to the biological source material or hybridization on its tab in the experiment.

If you try to delete a biomaterial or hybridization that is associated to another component and the deletion fails, caArray informs you of such, and directs you to other components that need to be deleted first.

Downloading Associated Data Files

caArray allows you to associate data files to sources, sample, extract, labeled extract or hybridization, as you import them into the system. You can also designate for caArray to auto-generate sources, samples, extracts, labeled extracts or hybridizations as you import appropriate files into the system. For more information, see *Importing Data* on page 86.

All imported data files that have been associated with a source, sample, extract, labeled extract or hybridization are listed in the **Download Data** section of its tab in an experiment (*Figure 4.12*).

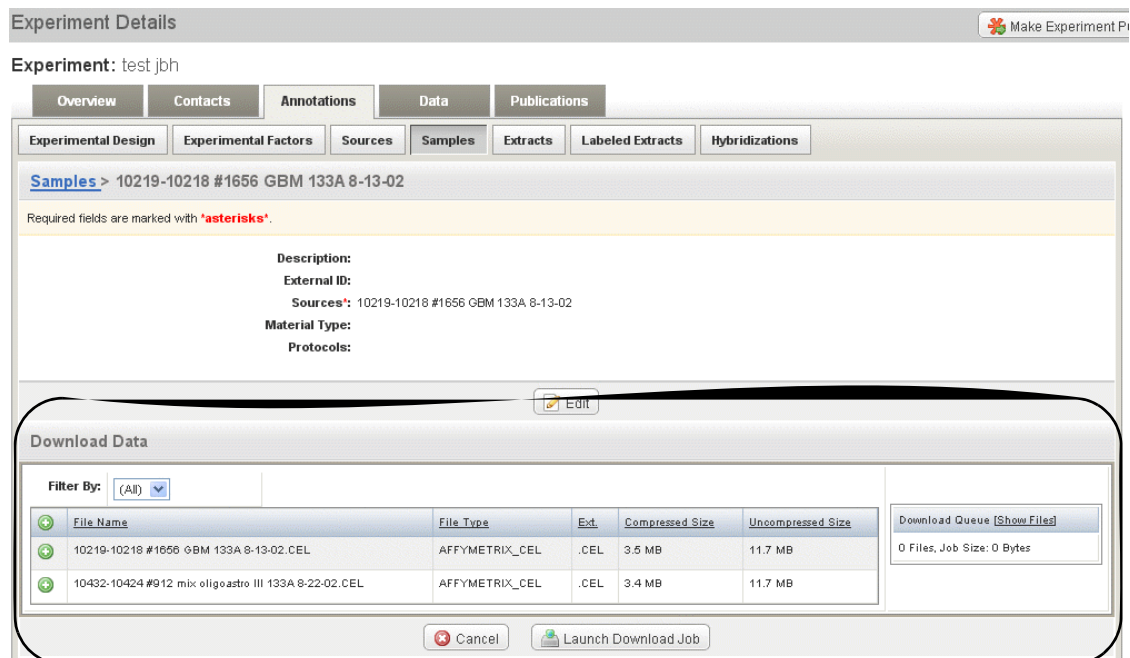


Figure 4.12 Files associated with biomaterials and hybridizations can be downloaded directly from the corresponding tabs.

Each file or subset of files can be selected for download.

1. To download data files associated with the source, click the green plus button (+) in front of each file you want to download
2. Click the **Launch Download Job** button.

Note: Clicking the **Launch Download** link downloads as a .zip file the data files (e.g. .CEL, .CHP, etc.) associated with this biomaterial or hybridization during the import process. Note that you can also download large data sets in batches. For more information, see *Importing Data* on page 86 and *Downloading Data from caArray* on page 93.

3. In the dialog box that opens, select to open or save the file to a disk.

Adding Vocabulary for Experiments

For an experiment you are creating, a number of the annotations subtabs (Experimental Design, Source, Samples, etc.) include panels where you can select or

add new attribute vocabulary terms to define or annotate experiment components.

Figure 4.13 The following figure displays an example page for performing this task.

The figure shows a web interface for selecting vocabulary terms. It consists of four main sections, each with a list of terms and a 'Selected' panel on the right.

- Tissue Site:** Filter: [], Add [+]. List: ALPHA2_VALUE (caArray), Bladder (caArray), Blood (DB:NCI_Thesaurus), Blood (EVS), Bone marrow (DB:NCI_Thesaurus). Selected Tissue Site: []
- Material Type:** Filter: [], Add [+]. List: ALPHA2_STAGE (caArray), ALPHA4_STAGE (caArray), ALPHA4_STAGE (CTO), brain (MO), Cell (NCI_Thesaurus). Selected Material Type: []
- Cell Type:** Filter: [], Add [+]. List: alpha2 (ArrayExpress), alpha2_cells (ALPHA2_STAGE), ALPHA4_STAGE (CAARRAY2.0), astrocyte (NCI_Thesaurus), B-Lymphocyte (NCI_Thesaurus). Selected Cell Type: []
- Disease State:** Filter: [], Add [+]. List: acute myeloid leukemia (DB:NCI_Thesaurus), acute myeloid leukemia (NCI_Thesaurus), Adenocarcinoma (DB:NCI_Thesaurus), ALPHA2_CONDITION (The Broad Institute of MIT and Harvard). Selected Disease State: []

Figure 4.13 Annotation panels for selecting or adding new vocabulary terms to experiments

You can enter terms for any of these attributes in three ways:

1. If the value for the attribute or condition displays in the site list, click the adjoining icon (+) to move it into the **Selected {attribute}** panel.
2. If the appropriate value is not displayed, to find a term of interest that might already be in the caArray dictionary, begin typing a term in the **Filter** text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.
3. To add a new term, click the **Add** button (+ Add). This takes you to the Manage {Attribute or Condition} page where you can add a new term. [Table 4.11](#) describes fields for defining the vocabulary term.

Vocabulary Term Category	Description of Fields
Term	
Value*	Enter the new term. <i>Example:</i> DNA
Description	Enter the description of the term, as appropriate. <i>Example:</i> deoxyribonucleic acid
Source	

Table 4.11 Fields for entering a new vocabulary term

Vocabulary Term Category	Description of Fields
Create a New Source [for the Term you are adding]	Select Yes or No <ul style="list-style-type: none"> If No, select from the drop-down list in the next field, the source for the term. In many cases, the source will be an existing controlled vocabulary such as the NCI Thesaurus, or the MGED Ontology (MO). If Yes, the dialog box expands with new fields where you can add the name, URL and version for the new source.
Source*	Select from the drop-down menu the source for the new term you are adding. This field disappears if you select Yes in the previous field.
Accession	
Accession URL	Enter the exact URL for accessing the new term. <i>Example:</i> http://mged.sourceforge.net/ontologies/MGEDontology.php#DNA
Accession Value	Enter the value given the term in the source vocabulary. <i>Example:</i> MO_945

Table 4.11 Fields for entering a new vocabulary term

- Once you have entered the appropriate information, click **Save**.

This returns you to the original tab, where you can continue defining the experiment attribute.

These same vocabulary management pages can also be accessed by clicking on Manage Vocabularies on in the left panel on the page. For more information see *Managing [Controlled] Vocabulary [Terms]* on page 72.

Data Tab

The Data tab is the location for uploading, validating, importing and downloading data relating to caArray experiments. When you click on the **Data** tab, four subtabs where you initiate data-related tasks display (*Figure 4.14*). They are described in *Table 4.12*.

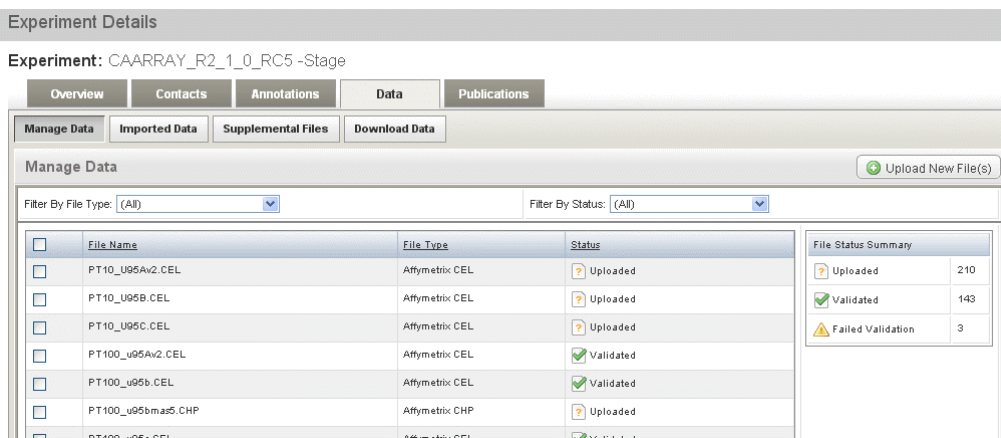


Figure 4.14 caArray experiment data tab, displaying several files that have been uploaded. A status summary displays for all files that have been uploaded into the experiment.

Data Tabs	Description
Manage Data	From this tab, you can perform data-related tasks such as uploading, validating and importing data into caArray. Additional tasks such as changing data file types and designating supplemental files also takes place here.
Imported Data	This subtab list all files that have been imported into caArray.
Supplemental Files	This tab lists files and documents that have been uploaded to caArray and have been designated supplemental on the Manage Data subtab.
Download Data	From this tab, you can download data that has been imported into caArray. If you are the owner of the experiment, uploaded data may also be download here.

Table 4.12 Tabs for performing data-related tasks

Note: To import data, you must have Write access to the experiment.

All of these and other data-related tasks are described in detail in [Chapter 6 Submitting Data to an Experiment](#) and [Chapter 7 Extracting Data from caArray](#).

Publications Tab

The Publications tab allows you to create, edit or delete associations of publications with the experiment you are creating or editing. This feature allows you to add journal articles, reviews, and books.

To add publication associations, follow these steps:

1. Locate and open the experiment for which you want to associate publications.
2. Click the **Publications** tab.
3. Click the **Add a New Publication** button.
4. On the form that opens, enter the appropriate information for the article or publication in the fields provided (and described in [Table 4.13](#)).¹⁰

Publications Fields	Description
Title*	Title of the publication
Authors	Author(s) of the publication
URL	URL for locating the publication
Editor	Editor of the publication, where appropriate
Pages	The page number(s) of the article your are referencing, where appropriate,
Publisher	Publisher of the publication

Table 4.13 Fields for documenting Publications

10. Fields with a red asterisk * are required.

<i>Publications Fields</i>	<i>Description</i>
PubMedID	ID for locating the publication in PubMed
Volume	Volume where article is found
Year	Year of publication
Publication	Name of the publication where article is found
Type	Select in the drop-down menu the publication type.
Status	Select in the drop-down menu the publish status: Published, In Preparation, Submitted, In Print.

Table 4.13 Fields for documenting Publications

5. After entering the appropriate information, click **Save**. A message displays verifying that the publication was successfully saved with the experiment.

The System saves the experiment with the associated publications and returns you to your experiment workspace, which is now updated with the state of the project.

Experiment Status Settings

As you enter details of the experiment on the Overview tab, you must click the **Save** button to open more tabs for recording experiment information. As you step through the successive tabs, you should continue save the experiment information by clicking the **Save** button. This saves the experiment in draft status, which means that the experiment cannot yet be shared with collaborators. By default, experiments with Draft status are not visible to Public users. This visibility can be changed on the Manage Permissions page of that experiment, as noted in the second bullet below.

- On each tab, you can “submit” the experiment by clicking the **Submit Experiment Proposal** button in the upper right corner, thus converting the experiment status to “In Progress”.
- An experiment in progress can be converted to public status by clicking the **Make Experiment Public** button in the upper right corner.
- When an experiment is public, you can click the **Retract Public Experiment** button after which the data associated with that experiment will no longer be downloadable by Public users.

Summary of experiment status and buttons to achieve that status:

Draft—Click **Save** on experiment tabs. Details of the experiment are available only to experiment owner (creator).

in Progress—Click the **Submit Experiment Proposal** button of an experiment in “draft” status. All details of the experiment or only selected segments of an experiment can be made available to all users or only to collaboration group(s) as “read-only, read-selective, read-write, or read-write selective”.

Public—Click the **Make Experiment Public** button of an experiment in “In Progress” status. All details of an experiment are then available to all users, including anonymous users.

Retract Public Experiment—Click this button to cancel the public status for an experiment. At that point, data associated with the experiment will no longer be viewable or downloadable by users other than the Experiment owner and designated collaborators.

For information about setting the visibility of an experiment, see *Experiment Visibility* on page 56.

Updating An Experiment Proposal


At any point, after you have saved an experiment draft, you, as its creator, can open the draft and edit it, add or delete data. These tasks can also be performed by collaborators with the appropriate permissions, when the experiment is In Progress.

Note: After data has been imported into an experiment, the array design associated with that data cannot be removed from the experiment's list of array designs.

Editing an Experiment

At any point after an experiment has been saved, as the experiment creator, you can edit it. From the Edit page, you can associate or edit protocols, edit or add annotations, or any other components for an experiment.

To edit an experiment, follow these steps:

1. In the My Experiment Workspace, locate the experiment you want to edit.
2. On the row corresponding to the experiment you want to edit, click the **Edit** button () and edit the data. You can also initiate an edit by clicking the Edit button at the bottom center of the page of an open experiment.

All information is editable except the automatically generated experiment ID and the status.

3. Click **Save** to save the edits to the draft.

Deleting an Experiment

An experiment can be deleted by the creator, as long as it is “Draft” status. To delete an experiment, follow these steps:


1. In the My Experiment Workspace, locate the experiment of interest on the **Work Queue** tab.
2. Click the **Delete** button that corresponds to the experiment.

Experiment Visibility

Once you create an experiment draft, it is listed on the **Work Queue** tab of your My Experiment Workspace. By default, draft experiments have no visibility to the public. As the experiment creator, once the experiment is In Progress (status indicated after you have “submitted” the experiment), you can change collaborator permissions. You can configure public access to the experiment, as well as collaboration group access.

Setting Public Visibility

To assign or modify experiment visibility, follow these steps:

1. Go to My Experiment Workspace and locate the experiment of interest on the Work Queue tab.
2. Click the **Permissions** icon () in the row listing the experiment.
3. The Experiment Permissions page that opens displays a panel that shows the current permission of the Experiment as assigned to the public and/or to all collaborator groups. It also describes the visibility policies of caArray and allows you to set visibility for your experiment (*Figure 4.15*).

Experiment: TCGA Analysis of RNA Expression for Glioblastoma Multiforme ([Details](#))

Assigning an Access Profile to this experiment allows you to control who gets access. Assign specific access rules down to the sample level.

Who May Access this Experiment	Control Access to Specific Content for	Experiment Access
The Public <input type="button" value="Edit Access Control"/> Edit access to the experiment by anonymous users The following access profiles can be managed to grant visibility into the summary and/or contents of an experiment prior to its publication. The profiles are: Visible - applied by default to each experiment upon creation, this profile exposes summary information without access to annotation and array data. Read - grants read access to the experiment as a whole - providing a preview into its content Read Selective - grants selective access to specific sample annotation and data No Visibility - completely removes the experiment summary information from view <input type="checkbox"/> Use TCGA Policy (applies only to Human SNP data) <input type="button" value="Save"/>	Collaboration Group: TCGA-DCC The Public	Read/Write Read

Collaboration Groups
 Edit access to the experiment by members of selected group
 Collaboration Group: 11/18/2008

Figure 4.15 Experiment Permissions page

Click the **Edit Access Control** button in the Public section to set the visibility for the public. In the right panel, select from the public visibility options, described as follows:

- **Visible**--applied by default to each experiment upon creation, this profile exposes summary information without access to annotation and array data
- **Read**--grants read access to the experiment as a whole - providing a preview into its content
- **Read Selective**--grants selective access to specific sample annotation and data. Select this option to apply selective access to experiment samples only. For more information, see *Setting Selective Permissions* on page 59.
- **No Visibility**--completely removes the experiment summary information from view

Note: Your ability to set visibility options varies according to the draft/ submission status of the experiment. See below.

The “TCGA Policy” option at the bottom of the left hand panel relates to The Cancer Genome Atlas (TCGA) policy for human SNP data, which is that only four fields of annotation can be provided to the open public:

- Clinical Diagnosis
- Histologic Diagnosis
- Tissue Anatomic Site
- Pathologic Status

Click this option to assign TCGA policy limits to your experiment.

For an Experiment in Draft status:

Note: Only **Visible** and **No Visibility** options are available for an experiment in draft status. The **Collaboration Groups** visibility option is not available to an experiment in Draft status.

- a. To control public access, under **The Public**, click the **Edit Access Control** button.
- b. In the Control Access to Specific Content panel that displays on the right, select the visibility option from the drop-down list. The only available options for an experiment draft) are: **Visible** and **No Visibility**.
- c. Click **Save** under the TCGA option to execute the Public access choices.

For an Experiment “In Progress”:

Note: Four visibility options plus collaboration group visibility are available for experiments that have been submitted to the caArray repository.

- a. To control public access, under **The Public**, click the **Edit Access Control** button.
- b. In the Control Access to Specific Content panel that displays on the right, select the visibility option from the drop-down list. The four available options are those described above: **Visible**, **Read**, **Read Selective** and **No Visibility**.

Note: **Read Selective** provides the option to assign visibility to just samples of an experiment. For more information see *Setting Selective Permissions* on page 59.

- c. The option at the bottom of the left hand panel relates to The Cancer Genome Atlas (TCGA) policy (for human SNP data), which is that only four fields of annotation can be provided to the open public:
 - Clinical Diagnosis
 - Histologic Diagnosis
 - Tissue Anatomic Site
 - Pathologic Status


Click this option to assign TCGA policy limits to your experiment.

- d. Click **Save** under the TCGA option to execute the Public access choices.

Setting Selective Permissions

caArray provides the option for you to selectively assign permissions to samples within an experiment that may have different visibility. This feature is available only in experiment with an “In Progress” status.

To assign selective permissions, follow these steps:

1. In your My Experiment Workspace, locate the experiment of interest on the Work Queue tab.
1. Click the **Permissions** icon () in the row listing the experiment.
2. Click the **Edit Access Control** button at the top of the dialog box, in the Public section. This opens the Control Access for Specific Content... dialog box (circled in [Figure 4.16](#)).

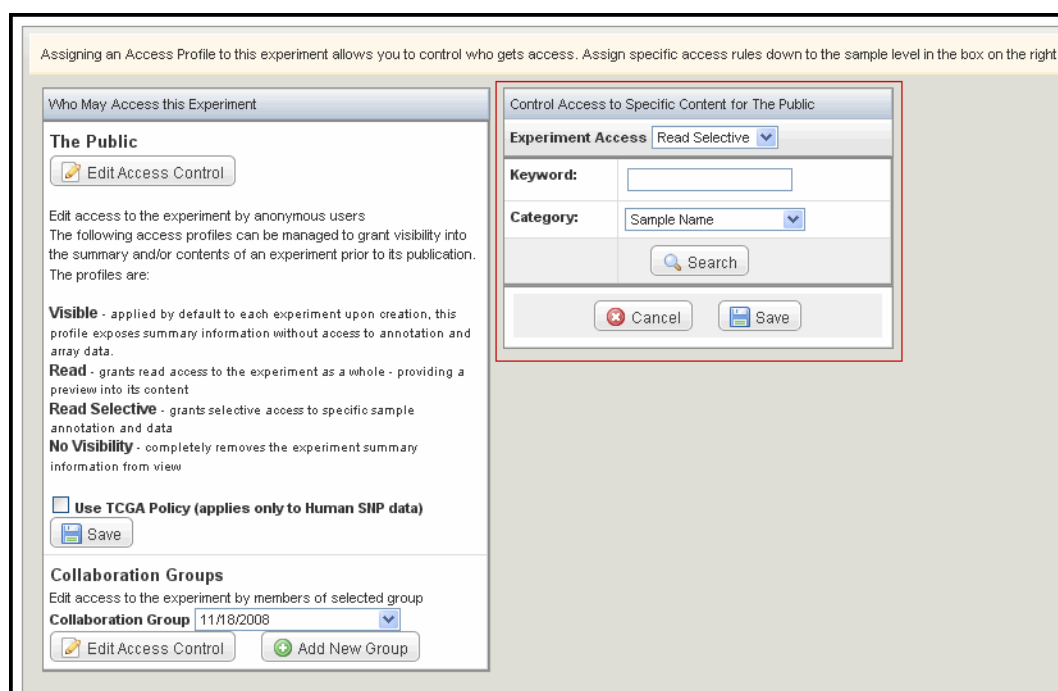


Figure 4.16 Permissions page with Read Selective Access dialog box on the right

3. Click the **Experiment Access** down arrow. This opens the list of visibility options.

- Click the **Read Selective** option. This expands the dialog box, showing additional options. From here you can search for the sample(s) to which you want to selective assign visibility (*Figure 4.17*).

Figure 4.17 Dialog box for selecting/searching for sample to which selective visibility can be assigned


- Enter a search keyword in the **Keyword** field.
- Click the down arrow in the **Category** field, and select the sample metadata criteria for the search.
- Click the **Search** button to launch the search.
- Select the samples you want from the returned search results, and change privileges for them by selecting the appropriate visibility option in the list box (**None**, **Read**, etc.).

Note: The **Edit Access Control** button in the Collaboration Groups section of the page works in the same way, except visibility is applied to Collaboration Groups, not just individual users.

Setting Collaboration Group Visibility

A section at the bottom of the left panel of the Experiment Permissions page (*Figure 4.15*) allows you to set experiment visibility for collaboration groups. You can use one of the existing groups, or you can create a group from this page. For information about how to create collaboration groups through the user management features, see *Managing Collaboration Groups* on page 99.

To configure experiment visibility for a collaboration group, follow these steps:



- Go to My Experiment Workspace and locate the experiment of interest on the Work Queue tab.
- Click the **Permissions** icon () in the row listing the experiment for which you want to alter permissions.

The Experiment Permissions page that opens displays a panel that describes the visibility policies of caArray and allows you to set visibility for your experiment. For information about setting public visibility, see *Experiment Visibility* on page 56.

If the Collaboration Group already exists:

1. At the bottom of the left-hand panel of the page, select the collaboration group of interest from the drop-down list.
2. Click the **Edit Access Control** button.
3. In the Control Access to Specific Content to [Group] panel that displays on the right, select the visibility option for the group from the Experiment Access drop-down list. The five available options are:
 - **None**--the collaboration group has no special privileges to access the experiment (apart from the privileges that are granted to a Public user).
 - **Read**--grants read access to the experiment as a whole, providing a preview into its content
 - **Read Selective**--grants selective access to specific sample annotation and data. For more information see *Setting Selective Permissions* on page 59.
 - **Read/Write**--grants read access to the experiment and the right to edit or otherwise modify the experiment
 - **Read/Write Selective**--grants selective read/write access to specific sample annotations and data
4. Click **Save** in that panel to execute the collaboration group visibility choices.

If the Collaboration Group must be created:

1. If the group to which you want to assign experiment visibility does not yet exist, click the **Add New Group** button.
2. In the Add a New Collaboration Group form, enter the name of the group and click **Save**.
3. The Manage Collaboration Groups page that opens display all groups in the system, including the one you just created. Click the **Edit** icon () to specify members for the group. For more information about creating and working with a collaboration group, see *Managing Collaboration Groups* on page 99.
4. Return to the Experiment Permissions page to continue assigning visibility for the experiment. You may need to return to the My Experiment Workspace and re-click the **Permissions** icon () in the row listing the experiment to get back to that page.
5. Continue from step 1 in this section of this topic: *If the Collaboration Group already exists*:

See also *Setting Public Visibility* on page 57.

CHAPTER 5 CURATION TOOLS

This chapter describes the processes for completing curation tasks in caArray.

Topics in this chapter include the following:

- *Managing Array Designs* on page 64
- *Managing Protocols* on page 68
- *Managing [Controlled] Vocabulary [Terms]* on page 72

Curation Tasks

Curation tasks in caArray 2.2 are available to all logged in users. These tasks govern global, and reusable data elements, namely array designs, protocols and vocabulary terms. Curation tasks are available under the Curation options on the left sidebar (*Figure 5.1*):

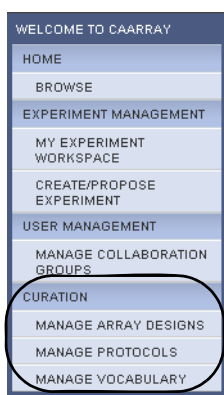


Figure 5.1 Curation options display in the left sidebar

Managing Array Designs

Any caArray user can upload, validate and import array designs for the supported providers shown in Table 5.1. caArray allows upload of a zip file containing multiple files for an array design.

Note: After data has been imported into an experiment, the array design associated with that data cannot be removed from the experiment's list of array designs.

<i>File Types</i>	<i>Validated and parsed as part of the import process</i>	<i>Imported but not parsed</i>
Array Design files	<ul style="list-style-type: none"> Affymetrix CDF, PGF CLF Illumina Design CSV Genepix GAL 	<ul style="list-style-type: none"> Agilent CSV, XML UCSF Spot SPT ImaGene TPL Nimblegen NDF MAGE-TAB ADF <p>Note: Any array design reference in a MAGE-TAB SDRF must refer to the LSID of an array design that has already been imported into caArray.</p>

Table 5.1 Array design file types that can be imported into caArray

Any caArray user can view, edit or replace the files. An array design only needs to be loaded once and is available to all users.

Viewing Array Designs

To view array designs in the system, follow these steps:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Array Designs**.

The Manage Array Designs page that opens displays all array designs that have been imported into caArray ([Figure 5.2](#)). See *Importing Data* on page 86. Properties corresponding to those array designs are described in Table 5.2 .

Manage Array Designs								
Array Designs								
43 items found, displaying 1 to 20. [First][Prev] 1, 2, 3[Next][Last] Import a New Array Design								
Array Design Name	Provider	Assay Type	Version Number	Feature Type	Organism	Edit	Edit File	Status
GenomeWideSNP_6.Full	Affymetrix	SNP	rev2	in_situ_oligo_features	Homo sapiens			
GenomeWideSNP_5.Full	Affymetrix	SNP	1	in_situ_oligo_features	Homo sapiens			
HG-U133_Plus_2	Affymetrix	Gene Expression	1	in_situ_oligo_features	Homo sapiens			
Mapping250K_Nsp	Affymetrix	Gene Expression	rev4	in_situ_oligo_features	Homo sapiens			
Mapping250K_Slv	Affymetrix	Gene Expression	rev4	in_situ_oligo_features	Homo sapiens			
HG-U133A_2	Affymetrix	Gene Expression	2	in_situ_oligo_features	Homo sapiens			

Figure 5.2 Array Designs imported into caArray

Note: Columns with underlined headings are sortable by clicking on the heading.


<u>Array Designs Properties</u>	Description
<u>Array Design Name</u>	Name assigned to the array design
<u>Assay Type</u>	<p>The assay type used for the Array Design.</p> <ul style="list-style-type: none"> • Gene Expression--experiment using microarrays intended to measure levels of transcribed genes • SNP--experiment using microarrays intended to detect nucleotide changes in chromosomal DNA • aCGH--array Comparative Genomic Hybridization; a method for the analysis of chromosome copy number changes (gains/losses). • Exon--Exon arrays are designed to study which exons are present in an expressed gene. • microRNA--Experiment that measures activity among the 217 genes encoding miRNA. Patterns of gene activity that can distinguish types of cancers can be discerned. • Methylation--experiment that attempts to establish patterns of methylation genome-wide or within targeted promoters or CpG islands
<u>Provider</u>	<p>Select from the drop-down menu the provider of the array.</p> <p>Note: Only Affymetrix, Illumina and GenePix formats are fully supported with validation and parsers in caArray 2.2. For more information, see the Note about File Types in <i>Managing Data</i> on page 77.</p>
<u>Version Number</u>	The version number of the array design
<u>Feature Type</u>	The technology type or platform of the reporters on the array. Note that these terms are from the MGED Ontology.
<u>Organism</u>	The organism the array was designed to assay.
Edit	If you do not have permissions to edit this Array Design, this icon is not visible. If it is, click the Edit icon () to open the Array Designs details page where you can edit the data. For more information, see the following section.
Status	Imported

Table 5.2 Array Designs properties

If you click the **Array Design Name** in the Import Array Designs page, the details page that opens displays the name including file type extension of the uploaded/imported array design file.

Adding an Array Design

To add an array design to caArray, follow these steps:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Array Designs**.

The Import Array Designs page that opens displays all array designs that have been imported into caArray (see previous topic [Importing Data](#)).

- To import a new array design, click **Import a New Array Design** in the upper right corner. This opens the Manage Array Designs dialog box.

The screenshot shows a web-based dialog box titled "Manage Array Designs" with a "Help" and "Print" link in the top right. Inside, there's a sub-dialog titled "New Array Design (Step 1)". Under the "Array Design Details" header, there are several input fields: "Description" (a large text area), "Assay Type*" (a dropdown menu), "Provider*" (a dropdown menu), "Version Number*" (a text input), "Feature Type*" (a dropdown menu), and "Organism*" (a dropdown menu). The asterisks on the dropdown menus indicate they are required fields. At the bottom of the sub-dialog are "Cancel" and "Next" buttons.

Figure 5.3 Manage Array Designs dialog box

- On the form that opens, enter the appropriate information in the Array Design Details fields provided (and described in Table 5.3).¹¹

Array Designs Details Properties	Description
[Array Design Name]	Note: This field does not appear in the array design form; the name is automatically generated by caArray based on the file you import.
Description	Enter an appropriate description for the array design you are adding.

Table 5.3 Array Designs properties

11. Fields with a red asterisk * are required.

Array Designs Details Properties	Description
Assay Type*	<p>The assay type used for the array design.</p> <ul style="list-style-type: none"> • Gene Expression--experiment using microarrays intended to measure levels of transcribed genes • SNP--experiment using microarrays intended to detect nucleotide changes in chromosomal DNA • aCGH--array <u>C</u>omparative <u>G</u>enomic <u>H</u>ybridization; a method for the analysis of chromosome copy number changes (gains/losses). • Exon--Exon arrays are designed to study which exons are present in an expressed gene. • microRNA--Experiment that measures activity among the 217 genes encoding miRNA. Patterns of gene activity that can distinguish types of cancers can be discerned. • Methylation--experiment that attempts to establish patterns of methylation genome-wide or within targeted promoters or CpG islands
Provider*	The provider of the array design. This is generally the company or group that manufactured the array design.
Version Number*	The version number of the array design
Feature Type*	The technology type or platform of the reporters on the array.
Organism*	The organism used for the Array described by the array design.

Table 5.3 Array Designs properties

4. Click the **Next** button.
5. In the **Upload Array Design File** section, click the **Browse** button to navigate to the file.
6. Select the **File Format** in the drop-down list. Compatible file types are listed at the beginning of this topic.

Note: If you select **Automatic**, caArray tries automatically to infer the array design type.

7. To add multiple files, click the **Add More Files** button, browse and select the additional file(s).
8. Click **Save** to launch the array design import process.

The process includes uploading the file, validating it and importing it into the system, all background processes. You should not leave the user interface once this process is underway, or you may have to start all over.


Note: After data has been imported into an experiment, the array design associated with that data cannot be removed from the experiment's list of array designs.

Editing an Array Design

In caArray 2.2, any logged in user can edit an array design. An array design already in the system can be replaced by a new array design by following the edit steps.


Note: After data has been imported into an experiment, the array design associated with that data cannot be removed from the experiment's list of array designs. Conversely, once an array design has been imported, the file associated with it cannot be changed.

To edit an array design, follow these steps:

1. On the row corresponding to the array design, click the **Edit** icon ().

OR

1. Open the array design by clicking on its name, and click the **Edit** button

( Edit) at the bottom of the details page.

2. All required fields become editable; enter any edits.
3. From this page, you can initiate uploading of a new array design file to replace the existing file.
4. Save any edit by clicking the **Save** button.

Deleting an Array Design

Array designs that are not associated with an experiment can now be deleted. To delete an array design, follow these steps:

1. On the row corresponding to the array design, click the **Delete** icon ().

Once you click the Delete icon, the file is physically deleted from caArray without requesting further confirmation.

Managing Protocols

In caArray, you can create and manage protocol(s) for referencing in an experiment. A protocol provides detailed documentation about the precise actions taken in any procedure that might be part of an experiment. For example, a protocol could describe the steps a laboratory used for any kind of process used in an experiment, such as the way a source material or sample is derived, the method used for labeling an extract or the methods used for running a hybridization or creating an image file of array results.

A protocol can be created independently of a specific experiment, or added during the course of creating biomaterials or a hybridization for an experiment. See the topics under *Annotations Tab* on page 34 for more information. A protocol can be used by any caArray user but it can only be modified by the owner of the protocol or another user with assigned permissions.

Viewing Protocols

To view existing protocols in caArray, follow these steps:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Protocols**. The **Protocols** tab displays by default (*Figure 5.4*).

Name	Type	Source	Description	Contact	Url	Edit
broad.mit.edu/hybridization/HT_HG-U133A-01	hybridization	Caarray	http://www.affymetrix.com/support/technical/index.affx			
broad.mit.edu/labeling/HT_HG-U133A-01	labeling	Caarray	http://www.affymetrix.com/support/technical/index.affx			
EXTPRTCL10654	nucleic_acid_extraction	Caarray	Approximately 10*6 cells were lysed in RLT buffer (Qiagen). Total RNA was extracted from the cell lysate using an RNeasy kit (Qiagen).			
EXTRACTION	nucleic_acid_extraction	Caarray	Lysates were captured with chloroform and purified using Qiagen RNeasy Mini Kit.			
GROWTH	grow	Caarray	Cell lines were plated in triplicate and lysed in Trizol.			
GROWTHPRTCL10653	grow	Caarray	TK6 cells were grown in suspension cultures in RPMI 1640 medium supplemented with 10% horse serum (Invitrogen, Karlsruhe, Germany). The cells were routinely maintained at 37 C and 5% CO2.			
HYBRIDIZATION	hybridization	Caarray	Hybridization was performed according to the manufacturer's protocol			
Hybridization-EukGE-V52v5	unknown_protocol_type	Caarray				
LABELING	labeling	Caarray	cDNA was prepared from 5 ug total RNA using the Invitrogen SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Inc, Carlsbad, CA) and amplified using the ENZO BioArray High-Yield RNA Transcript Labeling Kit (Enzo Biochem, Inc. New York, NY).			
P-AFFY-2	unknown_protocol_type	ArrayExpress				

Figure 5.4 Protocols page

All protocols that have been created in caArray display on this tab. Properties corresponding to those protocols are described in Table 5.4.

Protocol Properties	Description
<u>Name</u>	Name assigned the protocol
<u>Type</u>	Descriptor of the protocol type, such as labeling or hybridization.
<u>Description</u>	Description of the protocol procedure. Include any and all appropriate details, such as the detailed steps taken in a laboratory procedure. <i>Example:</i> Enter a description of a procedure for labeling RNA with fluorescent tags to be used in a hybridization procedure.
<u>Contact</u>	The name of the person to contact for information about the protocol.
<u>URL</u>	Link to a source of external documentation related to the protocol
Edit	Click the Edit icon () to open the protocol details page where you can edit the data. For more information, see <i>Editing a Protocol</i> on page 71.

Table 5.4 Protocol properties

Note: Columns with underlined headings are sortable by clicking on the heading.

2. To view details of a protocol, click its name.

Note: Any protocols you did not create are in read-only mode.

Viewing Protocol Types

To view existing protocol types in caArray, follow these steps:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Protocols**.
2. By default, the Manage Protocols page displays. To view protocol types, click the **Protocol Types** tab.

The **Protocol Types** tab displays all protocol types that have been created in caArray. Properties corresponding to those protocol types are described in Table 5.4.


Protocol Type Properties	Description
Value	The descriptor of the protocol type, such as labeling or hybridization.
Description	The description of the protocol type.
Source	The controlled vocabulary that is the source for the descriptor term value for the protocol type. The source name is a hypertext link that takes you to the website for the source.
Edit	Click the Edit icon () to open the protocol type details page where you can edit the data. For more information, see <i>Editing a Protocol Type</i> on page 72.

Table 5.5 Protocol properties

3. To view details of a protocol type, click its **Value**.

Creating a Protocol

To create a protocol, follow these steps:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Protocols**.

The Manage Protocols page that opens displays all protocols that have been created in caArray (*Figure 5.4*).

2. To create a new protocol, click the **Add Protocols** button in the upper right hand corner of the page.
3. In the Manage Protocols form that opens, enter the appropriate information for the new protocol. Fields are described in Table 5.6.¹²

Protocol Properties	Description
Name*	Name assigned the array design

Table 5.6 Protocol fields

12. Items with an asterisk are required.


Protocol Properties	Description
Description	Description of the protocol procedure. Include any and all appropriate details, such as the detailed steps taken in a laboratory procedure. <i>Example:</i> Description of a procedure for labeling RNA with fluorescent tabs to be used in a hybridization procedure.
Type*	<p>Descriptor of the protocol type such as “labeling” or “hybridization” from a controlled vocabulary, for example MGED.</p> <p>If the appropriate protocol displays in the list, click the adjoining icon () to move it into the Selected Protocols panel. Note: The available selections are limited based on the protocol type selected above.</p> <p>If the appropriate value is not displayed, to find a protocol of interest that might already be in the caArray, begin typing a term in the Filter text box. The available terms that display in the panel below are limited according to the text you enter. A message displays if a corresponding term cannot be found.</p> <p>If the appropriate protocol has not been entered into the system, click Add to open the page where you can add a new protocol.</p>
Contact	The name of the person to contact for information about the protocol.
Software	Name of software used in the protocol. <i>Example:</i> GenePix Pro 3.0.1.22
Hardware	Name of hardware used in the protocol. <i>Example:</i> GeneChip [®] Fluidics Station 450 [®]
URL	Link to a source of external documentation related to the protocol


Table 5.6 Protocol fields

- Click **Save** to save the protocol. Click **Cancel** to halt the action. In both cases, you are returned to the Manage Protocols page. The protocol you just added is listed first in the list of protocols.


Editing a Protocol

Note: A protocol can be edited by anyone, not just the owner of the protocol. All experiments that reference the protocol will be updated to reflect the changes made.


To edit a protocol, follow these steps:

- After logging into caArray, on the left sidebar under **Curation**, click **Manage Protocols**.
- On the row corresponding to the protocol, click the **Edit** icon ().


OR

1. Open the protocol by clicking on its name, and click the **Edit** button () at the bottom of the details page that opens.
2. All information for a protocol is editable. Make the appropriate edits on the form that opens. The edit is performed using the same steps described in *Creating a Protocol* on page 70.
3. Save any edits by clicking the **Save** button. To abort the edit, click the **Cancel** button. This returns you to the Manage Protocols page.

Editing a Protocol Type

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Protocols**.
2. On the Manage Protocols page, click the **Protocol Types** tab.
3. On the row corresponding to the listed protocol type, click the **Edit** icon ().

OR

1. Open the protocol type by clicking on its name, and click the **Edit** button () at the bottom of the details page.
2. All information for a protocol type is editable. Make the appropriate edits on the form that opens.
3. Save any edits by clicking the **Save** button. To abort the edit, click the **Cancel** button. This returns you to the Manage Protocols page.

Managing [Controlled] Vocabulary [Terms]

In caArray, when you are creating or editing experiments, many experiment attributes are available for entering descriptive terms or annotations. These attributes are:

- **Tissue Site**
- **Cell Type**
- **Disease State**
- **Material Type**

See the following sections for information about working with vocabulary terms in caArray.

Viewing Vocabulary Terms

To view existing vocabulary terms in caArray, follow these steps:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Vocabulary**. The **Manage Vocabularies** page opens, displaying four tabs that correspond to the attribute vocabulary categories in caArray (*Figure 5.5*).



Figure 5.5 Manage Vocabularies page

All vocabulary terms that have been imported into caArray display on the tabs. Properties corresponding to those protocols are described in Table 5.4.


Protocol Properties	Description
<u>Value</u>	Name assigned the term
<u>Description</u>	Description of the vocabulary term
<u>Source</u>	The external source or standardized vocabulary that is the source of the term.
Edit	Click the Edit icon () to open the vocabulary details page where you can edit the information for the term. For more information, see <i>Editing a Vocabulary Term</i> on page 74.

Table 5.7 Protocol properties

Note: Columns with underlined headings are sortable by clicking on the heading.

2. To view details of a vocabulary term, click its value.

Adding Vocabulary Terms

In caArray, you can enter a new vocabulary term while you are adding annotations to an experiment (see *Adding Vocabulary for Experiments* on page 51) or you can work with vocabulary terms using one of the curation tools of the application, described in this section.

To enter a vocabulary term, follow these steps:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Vocabulary**.
2. Select the attribute category that corresponds to the term you want to add.
3. Click the Add {attribute} button on the upper right of the page.

This takes you to the Manage {Attribute or Condition} page where you can add a new term. Table 5.8 describes fields for defining the vocabulary term.

Vocabulary Term Category	Description of Fields
Term	
Value*	Enter the new term. <i>Example:</i> DNA
Description	Enter the description of the term, as appropriate. <i>Example:</i> deoxyribonucleic acid
Source	
Create a New Source [for the Term you are adding]	Select Yes or No <ul style="list-style-type: none"> If No, select from the drop-down list in the next field, the source for the term. In many cases, the source will be an existing controlled vocabulary such as the NCI Thesaurus, or the MGED Ontology (MO). If Yes, the dialog box expands with new fields where you can add the name, URL and version for the new source.
Source*	Select from the drop-down menu the source for the new term you are adding. This field disappears if you select Yes in the previous field.
Accession	
Accession URL	Enter the exact URL for accessing the new term. <i>Example:</i> http://mged.sourceforge.net/ontologies/MGEDontology.php#DNA
Accession Value	Enter the value given the term in the source vocabulary. <i>Example:</i> MO_945

Table 5.8 Fields for entering a new vocabulary term


- Once you have entered the appropriate information, click **Save**. Click **Cancel** to halt the action.

This returns you to the original Manage Vocabularies {attribute} tab.


Editing a Vocabulary Term

Note: A vocabulary term can be edited by anyone.

To edit a vocabulary term, follow these steps:

- After logging into caArray, on the left sidebar under **Curation**, click **Manage Vocabulary**.
- Select the attribute tab where the term you want to edit is listed.
- On the row corresponding to the protocol, click the **Edit** icon ().

OR

4. Open the term details page by clicking on its value, and click the **Edit** button () at the bottom of the page.

5. Make the appropriate edits on the page, using the same steps as described in *Adding Vocabulary Terms* on page 73.

You can edit details of the term itself, its source, and the accession number for the term in the source database. All information is editable. For more information about the fields for defining the terms, see *Adding Vocabulary Terms* on page 73.

6. Save any edits by clicking the **Save** button. To abort the edit, click the **Cancel** button. This returns you to the Manage Vocabularies page.

CHAPTER 6

SUBMITTING DATA TO AN EXPERIMENT

This [sectionchapter](#) describes the processes for submitting data such as annotation and array content into caArray experiments.

The following topics are part of this [sectionchapter](#):

- [Managing Data on this page](#)
- [Uploading Data Files on page 81](#)
- [Validating Data Files on page 83](#)
- [Importing Data on page 86](#)
- [Supplemental Files on page 91](#)
- [Importing MAGE-TAB Data on page 89](#)
- [Downloading Files on page 92](#)

Managing Data

Note: Array designs are not imported through the tasks described in this section. For more information about working with array designs, see *Managing Array Designs* on page 64.

In caArray, you can submit data to an experiment by performing the following tasks on the Experiment Data tab:

1. **Uploading data:** moving data into caArray from a local or networked location
2. **Validating data:** determining by caArray if the data uploaded is valid according to pre-determined rules
3. **Importing data:** making validated data available through the user interface and or an API for download from caArray.

About File Types in caArray

caArray supports the ability to upload, validate, parse and import many data file types from many providers. [Table 6.1](#) indicates the file types that caArray currently supports with full validation and parsing as well as those that can be imported without validation and parsing. For those files, validation and parsing is turned off and the end state of those files is “imported not parsed”. This allows for the system to recognize that those files need to be parsed as new parsers are developed.

For unrecognized files, there is an option to set the file type to **Supplemental** which turns off validation and parsing and allows for the file to be downloaded. For more information, see *Supplemental Files* on page 91.

The file types listed in [Table 6.1](#) **the following table** can be imported into caArray:

File Types	Imported after validation and processing	Imported without validation and parsing
Raw/processed data files	<ul style="list-style-type: none"> Affymetrix CEL, CHP GenePix GPR* Illumina CSV <p>*For GenePix .GPR files, the sample names are already implicit in the data files themselves. If such files are imported as part of a set of files including a MAGE-TAB SDRF, the SDRF file must contain all the sample names that are implicit. Otherwise, a validation error occurs.</p>	<ul style="list-style-type: none"> Affymetrix DAT, RPF, TXT, and EXP Agilent TSV, TXT Illumina IDAT, TXT ImaGene TIF, TXT Nimblegen GFF, TXT ScanArray CSV GEO SOFT GEO GSM MAGE-TAB Data Matrix
Array Design files	<ul style="list-style-type: none"> Affymetrix CDF, PGF CLF Illumina Design CSV Genepix GAL <p>Note: These can be uploaded, validated and imported only through the Manage Array Design feature described in <i>Managing Array Designs</i> on page 64.</p>	<ul style="list-style-type: none"> Agilent CSV, XML UCSF Spot SPT ImaGene TPL Nimblegen NDF MAGE-TAB ADF <p>Note: These can be uploaded, and imported only through the Manage Array Design feature described in <i>Managing Array Designs</i> on page 64.</p> <ul style="list-style-type: none"> MAGE-TAB ADF (Array Design Format) <p>Note: Any array design reference in a MAGE-TAB SDRF must refer to the LSID of an array design that has already been imported into caArray.</p>

Table 6.1 File types that can be imported into caArray



File Types	Imported after validation and processing	Imported without validation and parsing
MAGE-TAB files	<ul style="list-style-type: none"> • MAGE-TAB with single SDRF (Sample and Data Relationship Format) • MAGE-TAB  multiple SDRFs • IDF (Investigation Description Format) only, no referenced SDRFs <p>Note: Only one IDF is allowed per import, since the import is in the context of a single experiment.</p> <ul style="list-style-type: none"> • SDRF with only source > Hybridizations • ADF (Array Design Format) • MAGE-TAB with existing samples in caArray. 	<ul style="list-style-type: none"> • 

Table 6.1 File types that can be imported into caArray

Note: Image files cannot be validated or imported successfully into caArray 2.2. For more information, see *Image File Importing Issues in caArray* on page 105.

[Uploading Data Files](#)

[Validating Data Files](#)

[Importing Data](#)

[Steps for Importing Data](#)

[Supplemental Files](#)

Using the Data Tab

The Data tab ([Figure 6.1](#)) is the location for uploading, validating, importing and downloading data relating to caArray experiments. When you click on the **Data** tab for an experiment, four subtabs where you initiate data-related tasks display. They are described in [the following table Table 6.2](#).

Data Tabs	Description
Manage Data	From this tab, you can perform data-related tasks such as uploading, validating and importing data into caArray. During data import, you can also associate data files with biological source materials and hybridizations. Additional tasks such as changing data file types and designating supplemental files also takes place here.
Imported Data	This subtab list all files that have been imported into caArray.

Table 6.2 Tabs for performing data-related tasks

Data Tabs	Description
Supplemental Files	This tab lists files and documents that have been uploaded to caArray, and identified as supplementary (reference) files. Supplementary files are associated with an experiment as a whole and not with individual samples.
Download Data	From this tab, you can download data that has been imported into caArray.

Table 6.2 Tabs for performing data-related tasks

Additionally, the Data tab displays a file status summary for the experiment. For example, it lists the number of uploaded, validated, validation failures and imported files, and so forth.

Figure 6.1 Data tab displays much information about uploaded files

From the Data tab, you can filter the files by file type and file status. Select either filter drop-down list and select the file type/status you want to display.

All kinds of data can be uploaded and otherwise managed in caArray, although the majority of data will likely be annotation and array content files.

- Notes:**
- Importing MAGE-TAB is the only mechanism for entering annotations that are not displayed as generically available and editable fields in the annotation user interface. The unique data will be visible but un-editable.
 - Importing array design files is performed through the Curation tool in caArray, not on the Data tab. For more information, see *Managing Array Designs* on page 64.

Note: It is not possible to import source or sample annotations directly into caArray 2.2 from their respective tabs in the user interface. You can, however, import MAGE-TAB files that contain source, sample and other biomaterial information. See *Importing Data* on page 86. You can download data files from the biomaterials tabs, which indicate associations with sources, samples, extracts, labeled extracts and/or hybridization.

Uploading Data Files

[Validating Data Files](#)[Importing Data](#)[Supplemental Files](#)[Downloading Files](#)

Uploading Data Files

Through the process of uploading annotation and array data, the content becomes available for validation and import into caArray. Imported files can be shared for download or deleted.

- Notes:**
- Individual files or multiple files packaged in a .zip format can be uploaded.
 - An individual file (whether standalone or packaged in a zip) can only be half as large as the amount of memory on your server in order for validation and import processing to occur. caArray 2.2 has been tested using a 2GB allocation of memory and therefore the maximum size for any individual, un-zipped file is 1GB.
 - caArray supports the upload of .zip compressed files only. NO other compression formats are supported in v.2.2 for extraction. caArray automatically extracts the files from a .zip file, discards the original and displays each of the files in the .zip to the user, indicating that each has been uploaded.

[Steps for Uploading Files](#)[Deleting a File](#)[Validating Data Files](#)[Importing Data](#)[Importing MAGE-TAB Data](#)[Supplemental Files](#)

Steps for Uploading Files

Note: You can upload data only for experiments for which you have appropriate permissions. See *Experiment Visibility* on page 56.

Note: In caArray 2.2, the ability to upload files is restricted to files (raw or compressed in .zip format) that are 2GB or less. In addition, uncompressed files can only be half as large as the amount of memory on your server in order for validation and import processing to occur. caArray 2.2 has been tested using a 2GB allocation of memory and therefore the maximum size for any individual un-zipped file is 1GB.



To upload data into caArray, follow these steps:

1. Go to My Experiment Workspace and locate the experiment of interest on the Work Queue tab.
2. Click the Experiment ID for the selected experiment. This opens the experiment to the Overview tab.

3. Select the **Data** tab, and **Manage Data** subtab.
4. Click the **Upload New Files** button. This opens the Experiment Data Upload dialog box ([Figure 6.2](#)).

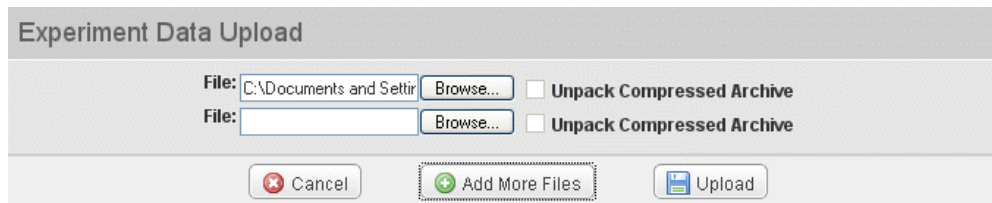


Figure 6.2 Experiment Data Upload dialog box

5. Click **Browse** to navigate to the file you want to upload.
 - If the file is compressed, check **Unpack Compressed Archive**.
 - To select several files for upload at once, click the **Add More Files** button. This opens more File/Browse options where you can locate several files.
6. Click the **Upload** button. Click **Cancel** if you decide to halt the task.

Note: When a zip of data files is being uploaded, if the zip file contains a directory, it will not upload, and an error message displays.

caArray launches the upload process. The process occurs in the background, allowing you to navigate through and use the application while the upload is in progress.

The Experiment Data Upload window, which *must remain open during the process*, monitors the percentage of the upload completed as well as its status. caArray will inform you when the upload process is complete.

When the upload has finished, click to close the window. The list of files displays on the Manage Data tab, as well as their status (uploaded) and file type. As you continue to work with the data, their status updates on the page (**Uploading**, **Uploaded**, **Validating**, **Validated**, **Importing** and **Imported**) ([Figure 6.3](#)).

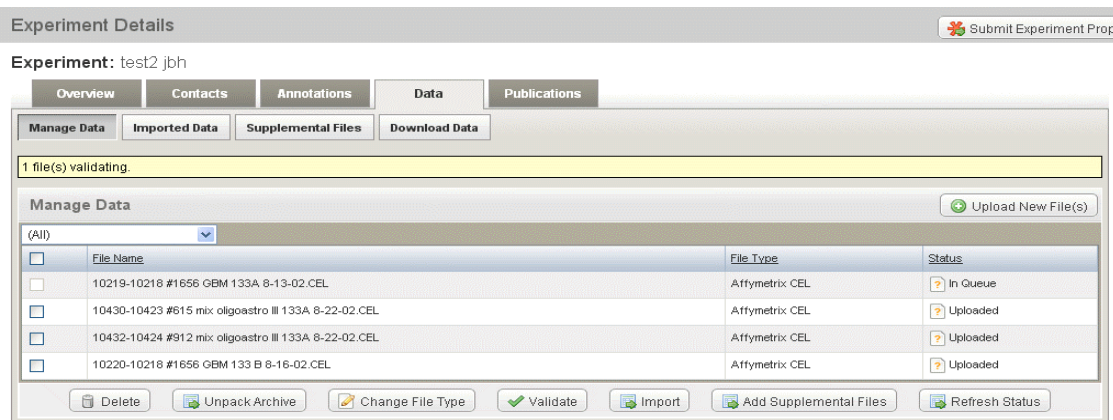


Figure 6.3 Manage Data subtab displays files in caArray and their status. The first listed file in "In Queue" to be validated.

Once files are uploaded, you should continue the files workflow by unpacking the archive, if necessary, then validating the file(s) and importing the file(s). Check boxes corresponding to each file allow you to select one or more at a time for further management. From the Manage Data tab, you can also change file types, designate files as supplementary, and delete files.

[Uploading Data Files](#)

[Deleting a File](#)

[Validating Data Files](#)

[Importing Data](#)

[Importing MAGE-TAB Data](#)

[Supplemental Files](#)

Deleting a File

On the Data tab, Manage Files subtab, check the box corresponding to one or more files you wish to discard and click the **Delete** button.

Note: Imported files can be deleted as long as they are not associated with any hybridization. For example, MAGE-TAB IDF and SDRF files can be deleted since they are not associated with any specific hybridization. Data files cannot be deleted once they are imported.

Once you click the **Delete** button, file(s) that meet the criteria (not associated with hybridizations) are physically deleted from caArray without requesting further confirmation.

[Uploading Data Files](#)

[Steps for Uploading Files](#)

[Validating Data Files](#)

[Importing Data](#)

[Importing MAGE-TAB Data](#)

[Supplemental Files](#)

Validating Data Files

Once data have been uploaded into caArray, anyone with “write” privileges for the experiment can validate and import annotation and array content files into the project. File validation verifies that data content adheres to a certain format; it does not evaluate the accuracy of the data from scientific viewpoint.

With proper permissions, you can upload the following file types which support information sharing. Note that because they are not array data files, they are not validated and no validation routines are available. You should identify these as Supplemental Files. Instructions are provided in *Supplemental Files* on page 91.

- Word documents
- Excel spreadsheets

- Power Point files
- PDFs

Supplemental files are not associated with any samples, but are associated at an experiment level.

In caArray, many file types can be uploaded, but not all file types can be validated (see definition above). All file types, even those that cannot be validated, can be imported. If you choose to import data that cannot be validated, validation is turned off, and a message indicating the data cannot be validated displays. The data gets imported, and its final state is “imported but not parsed”. For more information, see *Table 6.1* on page 78.

Uploading Data Files

Steps for Validating Data

Validation Errors

Importing Data


Importing MAGE-TAB Data

Supplemental Files

caArray Validation of MAGE-TAB

Steps for Validating Data

To validate uploaded data files in caArray, follow these steps:

1. Go to your My Experiment Workspace and locate the experiment of interest on the Work Queue tab.
2. Click the **Edit** button ( Edit). This opens the experiment to the Overview tab; you are in Edit mode.
3. Select the **Data** tab, and the **Manage Data** subtab.
4. Check the box corresponding to the file(s) you want to validate, and click the **Validate** button.

Note: If you choose to import a file before validating it, validation is launched automatically prior to import.

For files where the type cannot be inferred, their status is marked **Unknown**. You must change the file type to a known format before validation can proceed.

5. Upon launching validation, the status of the file on the Manage Data tab updates to **In Queue**, **Validating**, **Validated**, or **Validation Failed**.

The validation process takes place in the background, allowing you to continue to work in caArray, except in the file(s) being validated. You can check the status of the import by clicking the **Refresh Status** button at the bottom of the page.

caArray performs structural and then content validation against each file you have selected, updating the status of each file in the yellow message box periodically (10

seconds) until all files display the validation status in the **Status** column: **Validated** or **Validation Failed**.

[Uploading Data Files](#)

[Validating Data Files](#)

[Validation Errors](#)

[Importing Data](#)

[Importing MAGE-TAB Data](#)

[Supplemental Files](#)

Validation Errors

If validation fails, the file cannot be imported and a **Validation Failed** message displays on the Manage Data tab in the row corresponding to the file.

A validation error can be structural or content-based. Validation can fail for the following reasons:

- Format unknown (based on file extension and array type)
- Reference file not found
- File incomplete
- Vocabulary failure – annotation terms not found in supported ontology

To view a validation error description, click the hypertext **Failed Validation** link in the **Status** column

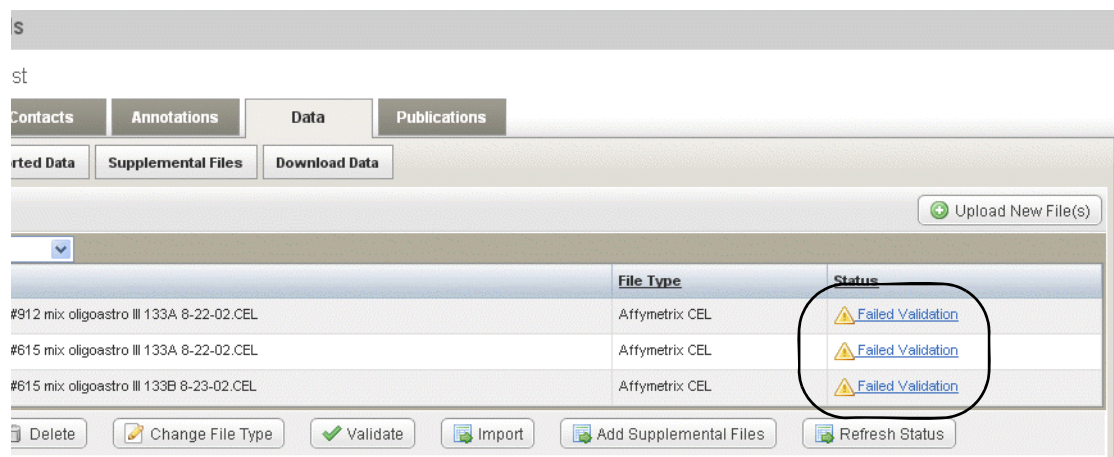


Figure 6.4 Validation failures display on the Manage Data table.

For more information, see [caArray Validation of MAGE-TAB](#).

[Uploading Data Files](#)

[Validating Data Files](#)

[Steps for Validating Data](#)

[Importing Data](#)

Supplemental Files

Importing Data

Note: This [section](#)[topic](#) describes the import of data files that are not MAGE-TAB files. It also introduces the procedure for importing all data files, including MAGE-TAB. For the specifics of importing MAGE-TAB files, see *Importing MAGE-TAB Data* on page 89 and [Appendix A, MAGE-TAB in caArray](#).

Once files are uploaded into caArray, only the person who uploaded the data can access it for validating and for importing annotation and array content files into the project. The Import feature allows the array data to be stored in the database, associated with the appropriate biomaterial and hybridization annotation. In addition, if a parser is available for this file type, discrete data values are available through the API. Data that has been imported is available to collaborators with the appropriate access and to the Public when the experiment is made public.

If you import just data files (for example, .cel, .chp, etc.), caArray automatically creates a source, sample, extract, labeled extract and hybridization for each data file. If multiple files with the same name (but different extensions) are imported, only one annotation chain of source > sample > extract > labeled extract > hybridization will be created, and all of the files will be associated with that single linked chain.

If you import a MAGE-TAB set (idf and sdrf) along with the data files, where the sdrf refers to each of the data files, the sdrf tells the system what sources, samples, extracts, labeled extracts and hybridizations to create. For more information, see *Importing MAGE-TAB Data* on page 89.

Files can be downloaded from the Data tab or from the Annotations tabs. For more information, see *Downloading Data from caArray* on page 93.

Note: While data is being imported into an experiment, all attributes and annotations of that experiment become read-only, so that user interface changes do not conflict with annotations being created as part of the import.

Uploading Data Files

Validating Data Files

About File Types in caArray


Steps for Importing Data

Importing MAGE-TAB Data

Supplemental Files

Steps for Importing Data

To import data, follow these steps:

1. Go to your My Experiment Workspace and locate the experiment of interest on the Work Queue tab.
2. On the row that corresponds to the experiment, click the **Edit** button (). This opens the experiment to the Overview tab in Edit mode.

3. Select the **Data** tab, and the **Manage Data** subtab.
4. Check the box corresponding to each file you want to import, and click the **Import** button.

Notes about import:

- If the file has not been previously validated, caArray performs the validation.
 - If validation fails, import does not proceed. Validation errors can be viewed as described in *Validating Data Files* on page 83.
 - If the validation completes successfully, caArray continues with the import and automatically auto-refreshes the status of the file set until the Import is complete and the Import Status of the file displays.
 - If some files in the set cannot be imported, caArray asks you if the remaining files should be imported. Click the appropriate answer.

5. In the Import Options dialog box that opens, you can associate selected files to existing biomaterials or annotations or specify for caArray to autocreate annotation sets of biomaterials. The details for each options is as follows:
 - **Autocreate annotation sets ... for each selected file:** If you select this option, for every unique file name that is imported, caArray creates a Source – Sample – Extract – LabeledExtract – Hybridization chain corresponding to each data file imported. These entities are identified based on the base name of the data file. For example, importing `mouse_342.txt`, `mouse_342.chp` and `mouse_342.cel` will result in one chain of biomaterials and hybridization, each named `mouse_342`.
 - **Autocreate a single annotation set ... for all selected files:** If you select this option, caArray creates a single Source - Sample - Extract - Labeled Extract - Hybridization chain, and associates all selected data files with this single chain.
 - **Associate selected file(s) to existing biomaterial or hybridization:** If you select this option, caArray displays all available sources, samples, extracts, labeled extracts and hybridizations. You select one of these, caArray associated the selected files with that biomaterial or hybridization. Note that additional items in the chain (to the right of the selected biomaterial) may need to be generated by the System.

Example: You import four data files: `zebrafish_6311.cel`, `zebrafish_6311.chp`, `zebrafish_6666.cel` and `zebrafish_6666.chp`. You choose to associate these data files with an existing extract called `zebrafish_extract_6`. caArray auto-generates a LabeledExtract – Hybridization chain called `zebrafish_6311` and associates the first two data files with it. caArray also auto-generates a LabeledExtract-Hybridization chain called `zebrafish_6666` and associates the last two data files with it. Both of these auto-generated chains will be associated with the `zebrafish_extract_6` Extract that you selected. The part of the chain to the left of the selected extract (sources, samples) will remain unmodified.

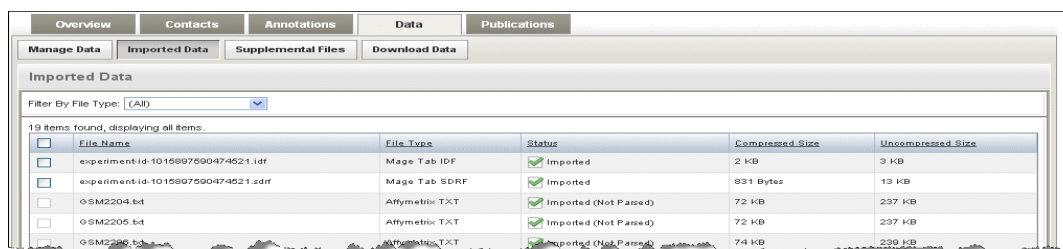
For more information regarding data import and its relationship to biomaterials, see [Importing MAGE-TAB Data on this page](#).

6. Select the appropriate option and click **Import**.

Once you launch the import, the status of the file on the Manage Data tab is updated. File status can be **In Queue**, **Validating**, **Validated**, **Importing**, **Imported**, **Imported not Parsed**, **Validation Failed** or **Import Failed**.

The import process takes place in the background, allowing you to continue to work in caArray. You can check the status of the import by clicking the **Refresh Status** button at the bottom of the page.

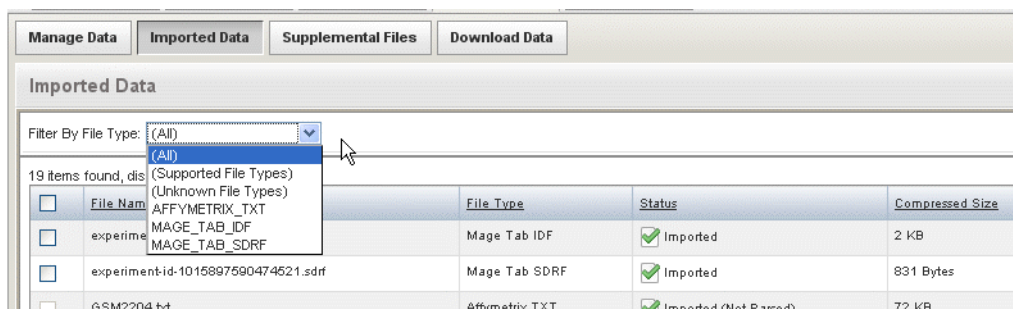
After a successful import, the files automatically move to the Imported Data subtab and the **Status** of the file set is **Imported** ([Figure 6.5](#)).



File Name	File Type	Status	Compressed Size	Uncompressed Size
experiment-id-1015897590474521.idf	Mage Tab IDF	Imported	2 KB	3 KB
experiment-id-1015897590474521.sdrf	Mage Tab SDRF	Imported	831 Bytes	13 KB
GSM2204.bt	Affymetrix TXT	Imported (Not Parsed)	72 KB	237 KB
GSM2205.bt	Affymetrix TXT	Imported (Not Parsed)	72 KB	237 KB
GSM2206.bt	Affymetrix TXT	Imported (Not Parsed)	74 KB	239 KB

Figure 6.5 Imported Data tab displaying imported data files

7. To filter the data display by file type, click the **Filter by File Type** drop-down arrow ([Figure 6.6](#)).



File Name	File Type	Status	Compressed Size
experiment-id-1015897590474521.idf	Mage Tab IDF	Imported	2 KB
experiment-id-1015897590474521.sdrf	Mage Tab SDRF	Imported	831 Bytes
GSM2204.bt	Affymetrix TXT	Imported (Not Parsed)	72 KB

Figure 6.6 Drop-down list for filtering imported files by file type

caArray creates the appropriate annotations (sources, labeled extracts, etc.) corresponding to the files imported.

[Uploading Data Files](#)

[Validating Data Files](#)

[Importing Data](#)

[About File Types in caArray](#)

[Importing MAGE-TAB Data](#)

[Deleting a File](#)

Importing MAGE-TAB Data

Before you can import MAGE-TAB data into caArray, you must first create the experiment into which the data is to be imported. For more information see *Creating an Experiment* on page 30.

This topic covers MAGE-TAB-specific details important for a successful data import into caArray. The import procedure itself is the same for all data files. For more information, see *Importing Data* on page 86.

For caArray-specific information regarding MAGE-TAB, see [Appendix A, MAGE-TAB in caArray](#).

See these additional topics:

- [MAGE-TAB in caArray--Overview](#)
- [Missing Biomaterials in MAGE-TAB are Auto-generated](#)

MAGE-TAB in caArray--Overview

MAGE-TAB is a simple spreadsheet-like format to capture annotation information for microarray experiments. It includes support for all the annotations required for MIAME compliance. The complete MAGE-TAB specification contains examples of experiments described using the MAGE-TAB format.

For more information, see the MAGE-TAB Specification document located here: <http://www.mged.org/mage-tab/spec1.0.html>. See also [Appendix A, MAGE-TAB in caArray](#) for details about how MAGE-TAB data is handled in caArray.

Array data can be uploaded, validated and imported as MAGE-TAB files. Experiment annotations can be captured using two types of MAGE-TAB files:

1. The Investigation Description Format (IDF) file, which provides high-level information about the experiment.

Warning: In a tab-separated line in the IDF, there must be no empty columns, i.e., two tabs with nothing between them. Empty columns will result in import failure.

2. The Sample and Data Relationship Format (SDRF) file that describes relationships between samples, arrays, data files, protocols, factor values, etc. The SDRF is a table in which each hybridization channel is represented by a row, and the columns represent the steps of the experiment, read from left to right.

In caArray 2.2, caArray does not parse the remaining two types of MAGE-TAB files:

3. Array Design Format (ADF) file
4. Data matrix file

Experiment data can be represented in native raw and derived files (for example, Affymetrix CEL, Affymetrix CHP, Illumina Group Probe Profile TXT). These data files are linked to the appropriate samples in the SDRF file.

Currently, caArray requires that the IDF, SDRF and all data files referenced in the SDRF be imported at the same time.

Note: If a data file being imported in the data set with the SDRF is not referenced in the SDRF, caArray generates a validation error and terminates the import process.

Procedures for uploading, validating and importing MAGE-TAB files are the same as for uploading other data file types in caArray.

See these additional topics:

- *Uploading Data Files* on page 81
- *Validating Data Files* on page 83
- *Importing Data* on page 86

MAGE-TAB Import with Existing Experimental Components

If the MAGE-TAB data set you are importing refers to biomaterials, hybridizations, data files or experimental factors that already exist in caArray, they will be reused. Only additive changes to linkages and nodes are allowed; you cannot delete existing linkages or nodes.

Biomaterial characteristics, material type, source provider and label attributes can all be modified. Does this mean during import? You can add new characteristics. During import, or after? Experimental factors and protocols cannot be added or changed for existing biomaterials or hybridizations. Attributes of existing persons, publications, experimental designs and factors cannot be changed.



Examples:

1. The user imports a set of MAGE-TAB plus data files, creating a Sample associated with 20 Extracts. The user later imports a new batch of MAGE-TAB plus data files, creating 20 new Extracts but associating them to the same Sample as in the first batch.
2. The user imports a set of MAGE-TAB plus raw data files. The user later modifies the SDRF to add derived data files (that were derived from the same raw data files that have already been imported), and imports this new SDRF with the derived data files.
3. The user imports a set of MAGE-TAB plus data files, creating a set of Samples and other annotations. At a later stage, clinical characteristics for those same Samples become available for the first time, or change from what was originally submitted. The user modifies the SDRF to add the new clinical characteristics (or to change values of already-existing characteristics), and re-imports the SDRF.



Missing Biomaterials in MAGE-TAB are Auto-generated

In the MAGE-TAB SDRF you are importing, if a biomaterial node is missing in the Source – Sample – Extract – Labeled Extract – Hybridization chain, appropriate intermediate nodes are automatically generated to complete the chain. The number of nodes generated will depend on the left side of the graph.

Examples:

1) If the SDRF describes one source connected to three extracts, one sample will be auto-generated and inserted in the chain. On the other hand, if the SDRF describes three sources combined to generate one extract, three samples will be auto-generated and inserted into the chain.

2) If the first (leftmost) biomaterial node in the SDRF is not a source, then the chain to the left of the biomaterial node will be auto-generated by caArray. For example, say the SDRF starts with the Extract column, and there are two unique extracts in the SDRF called “liver_2900” and “liver_3200”. caArray auto-generates two sources called “liver_2900” and “liver_3200”, and also two samples called “liver_2900” and “liver_3200”. The sources, samples and extracts are then linked in two separate chains.

If biomaterials missing in the SDRF are auto-generated, caArray “intelligently” associates protocol applications with the most appropriate auto-generated node. Note that this applies only to protocol applications in the SDRF whose association to biomaterial nodes is ambiguous.

Example:

If the SDRF contains a sample followed by a labeling protocol followed by a hybridization, the labeling protocol will be associated with the auto-generated extract > labeled extract portion of the chain. Similar logic is used for extraction and hybridization protocols.

Note: From the biomaterial (source, sample, extract, and/or labeled extract) or hybridization annotation pages (described in the sections beginning with *Biological Source Material* on page 37), you can download data files (e.g. .CEL, .CHP, etc.) that have been associated with these biomaterials and hybridizations. For more information, see *Downloading Data from caArray* on page 93.

[Uploading Data Files](#)

[Validating Data Files](#)

[Importing Data](#)

[About File Types in caArray](#)

[Supplemental Files](#)

Supplemental Files

Many file types can be uploaded into caArray, but only validated array content files can be imported into the application and parsed, therefore making the file content extractable through the API. Other files types can be designated on the Manage Data page as “supplemental files”.

To identify uploaded files as supplemental files, follow these steps:

1. Go to your My Experiment Workspace and locate the experiment of interest on the Work Queue tab.
2. Open the experiment and select the **Data** tab, and the **Manage Data** subtab.
3. Check one or more boxes for the file(s) you want to identify as supplemental files.

4. Click the **Change File Type** button.
5. On the drop-down list, for each appropriate file, scroll down and select **Supplemental File**. The file type then changes to Supplemental File.

- OR -

6. Change the file type in bulk by selecting all file types that need to be changed to a specific file type, and select the file type in the dialog box.
7. Back on the Manage Data subtab, make sure these files are still checked, and click the **Add Supplemental Files** button.

As you do so, the selected files move to the Supplemental Files tab.

Uploading Data Files

Validating Data Files

Importing Data

About File Types in caArray

Steps for Importing Data

Supplemental Files

Downloading Files

With appropriate permissions, you can extract files that have been imported into caArray. For more information, see [Chapter 7 Extracting Data from caArray](#).

CHAPTER 7

EXTRACTING DATA FROM CAARRAY

This chapter describes the processes for extracting data from the caArray repository.

Topics in this chapter include:

- [Downloading Data from caArray](#) on this page
- [Extracting Data Programmatically by API](#) on page 95

Downloading Data from caArray

Any user can download public data from any publicly available instance of caArray. No permission restrictions are required to acquire the data, either at an experiment level or sample level within or across projects.

From the biomaterial annotation pages associated with an experiment, you can download imported data files (e.g. .CEL, .CHP, etc.) that have been associated with those biomaterials or hybridizations. The files download as a .zip file. The MAGE-TAB files themselves must be downloaded from the Data tab described in this section. For more information about working with MAGE-TAB data, see *Importing MAGE-TAB Data* on page 89. For information about downloading data files from the Annotation tabs, see the sections beginning with *Biological Source Material* on page 37.

Note: Only files that have been imported into caArray can be downloaded on the biomaterial annotation pages. For more information, see *Importing Data* on page 86.

To download data of interest from the caArray Data tab, follow these steps:

1. Go to My Experiment Workspace and locate the experiment of interest on the Work Queue tab.
2. Open the experiment, and select the **Data** tab, and the **Download Data** subtab.

You can sort the columns of the list by clicking on the column headers. You can also filter the list of files by choosing the file type on the **Filter** drop-down list.

All files that are part of this experiment display on this tab (*Figure 7.1*).

The screenshot shows the 'Download Data' interface. At the top, there are two 'Filter By File Type' dropdown menus, both set to '(All)'. Below them is a table with columns: File Name, File Type, Ext, Compressed Size, and Uncompressed Size. The table lists 15 files, including 'experiment-id-1015897590474621.idf' and several 'GSM22' files. To the right of the table is a 'Download Queue' panel, which is circled in red. It contains a 'Show Files' link and displays '0 Files, Job Size: 0 Bytes'. At the bottom of the interface are three buttons: 'Clear Download Queue', 'Launch Download Job', and 'Export Consolidated MAGE-TAB'.

File Name	File Type	Ext	Compressed Size	Uncompressed Size
experiment-id-1015897590474621.idf	MAGE_TAB_IDF	.idf	2 KB	3 KB
experiment-id-1015897590474621.sdrf	MAGE_TAB_SDRF	.sdrf	831 Bytes	13 KB
GSM2215.bt	AFFYMETRIX_TXT	.bt	64 KB	224 KB
GSM2216.bt	AFFYMETRIX_TXT	.bt	64 KB	224 KB
GSM2217.bt	AFFYMETRIX_TXT	.bt	63 KB	222 KB
GSM2218.bt	AFFYMETRIX_TXT	.bt	63 KB	222 KB
GSM2219.bt	AFFYMETRIX_TXT	.bt	64 KB	224 KB
GSM2220.bt	AFFYMETRIX_TXT	.bt	64 KB	224 KB
GSM2204.bt	AFFYMETRIX_TXT	.bt	72 KB	237 KB
GSM2205.bt	AFFYMETRIX_TXT	.bt	72 KB	237 KB
GSM2206.bt	AFFYMETRIX_TXT	.bt	74 KB	239 KB
GSM2207.bt	AFFYMETRIX_TXT	.bt	73 KB	238 KB
GSM2208.bt	AFFYMETRIX_TXT	.bt	73 KB	238 KB
GSM2209.bt	AFFYMETRIX_TXT	.bt	73 KB	239 KB
GSM2210.bt	AFFYMETRIX_TXT	.bt	73 KB	238 KB

Figure 7.1 Download data subtab

3. To filter the list of files on the tab, click one of the following drop down menus:
 - a. Click **Filter by File Type** drop-down arrow to select the file types to be displayed.
 - b. Click the **Filter by File Status** drop-down arrow to select the files by their status to be displayed.
4. Identify one or more file(s) to be downloaded. Click the plus icon (⊕) to the left of the **File Name** column on the row of the file(s) you select. That places the data set in the download queue, visible in the right panel on the tab (circled). The identifying information for the file, including the total file size displays there.

Note: If you select multiple or all files to download, and there is a large amount of data, caArray calculates the total size of the download. If the size is greater than 1.5 GB (after compression), the system breaks up the download into batches of files, with each batch limited to a size not more than 1.5 GB (after compression). A download button is provided for each batch. When you have downloaded a batch, it is marked as having been downloaded.

5. To remove selected files from the queue, click the **Remove** icon(s) (⊖) corresponding to the data file or click the **Clear Download Queue** button.
6. You can generate MAGE-TAB files that consolidate all the annotation information about the experiment in the caArray repository. To perform this task, click the **Export Consolidated MAGE-TAB** button.

This generates an IDF and SDRF file containing information about the experiment including biomaterial-hybridization-data chains, biomaterial characteristics and material type, source providers, labeled extract labels, experiment title and description, and term source (vocabulary) information. Export of other attributes like experimental factors, protocols, publications, persons, etc. is deferred.

7. Click the **Launch Download Job** button to initiate the download process of files other than MAGE-TAB.

On launching the process, caArray displays the following message: *The Download job is being assembled*. The job proceeds until all designated files are downloaded.

- The length of time for the download is dependent upon the file size.
 - You can continue to work in caArray during the download process.
8. In the dialog box that opens, indicate whether you want to open or save the file to be downloaded. To save, navigate to the destination where the file will be saved, always with the title `caArray.zip`.

When the download is complete, your local system displays an on-screen message telling you that the download is finished or that it failed.

Grid Availability

For those Institutions that register their caArray instance on caGrid, the public data is available to the integrated tools that use the caGrid service. For more information, see <https://cabig.nci.nih.gov/workspaces/Architecture/caGrid>.

Note: The ability to browse and search from the caArray user interface features across the Grid is not available in caArray 2.2.

Extracting Data Programmatically by API

Data in caArray can be extracted programmatically using an API. caArray also provides a grid service which allows software engineers to acquire data from caArray.

For more information about using the remote Java API and grid service to extract data, see the *caArray 2.2 Technical Guide* which can be downloaded from this site: https://gforge.nci.nih.gov/frs/?group_id=305https://gforge.nci.nih.gov/frs/?group_id=305.

CHAPTER 8

USER ACCOUNT MANAGEMENT

This chapter describes the process for creating and managing accounts for users and groups of collaborators in caArray.

Note: The options for performing user management tasks are visible in caArray on the left sidebar of the browser only if you have these Admin privileges.

Topics in this chapter include:

- [Roles in caArray](#) on this page
- *Managing User Accounts* on page 98
- *Managing Collaboration Groups* on page 99

Roles in caArray

When a new user account is created within caArray, the user can be assigned one or more roles that then determines that user's access to specified private data.

Role	Description	Permissible Actions
Anonymous User	User without a caArray account or a non-logged in user Note: Because of the use of CSM, the Anonymous Group is crafted, but it does not need to be assigned by the System Administrator. caArray does it automatically upon Login.	Browse and Search tasks, downloading Public data
System Administrator	Person responsible for the effective operation of caArray	Manages users

Role	Description	Permissible Actions
Principal Investigator [PI]	Owns experiments and studies and/or projects	Manages experiments Manages experiment visibility
Lab Administrator	Responsible for managing lab operations. They typically interact with submitting investigators, assign work, and run reports on the operations of the lab.	Same as PI in caArray 2.2
Lab Scientist	The primary handler of samples in the lab. They run the experiments, collaborate with the statisticians and document their activities step by step.	Same as PI in caArray 2.2
Biostatistician	A special form of submitter who is responsible for statistical analysis of project data. The key actions to be performed are review of experiment designs, submission of quality control metadata, and uploading of normalized data and the annotation of the parameters used.	Same as PI in caArray 2.2
External System Note: Not listed under Roles	Systems other than caArray from which caArray data can be extracted programmatically using an API.	For more informations, see <i>Extracting Data Programmatically by API</i> on page 95.

When the account is registered and roles are assigned, the user can fully access caArray according to the roles provisioned.

Managing User Accounts

In caArray 2.2, all tasks related to creating and managing user accounts can be performed only by a System Administrator. The System Administrator must use the NCICB User Provisioning Tool (UPT) v. 3.2 for performing these tasks. All instructions for managing user accounts are described in the UPT 3.2 User's Guide which can be accessed from this website link to NCICB GForge: http://gforge.nci.nih.gov/frs/download.php/2634/UPT_User_Guide.pdf

Note: If you are a person who wants to be caArray user, you can launch a request for a user account from the Welcome/Login page. For more information, see *Requesting a User Account* on page 10

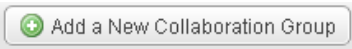
Managing Collaboration Groups

Any registered user in caArray can create, edit and delete collaboration groups and the users associated within them. This set of users (“collaborators”) can then be given access to an experiment for which the user is the Data Owner (usually the creator of the experiment) or to particular samples and their underlying array data. Only registered users are available to be a part of a collaboration group.

Creating a Collaboration Group

A group can be created by any registered user of caArray.

To create a group, as a logged in user, follow these steps:

1. Click the **Manage Collaboration Groups** option on the left sidebar.
2. On the Manage Collaboration Groups page, click the **Add a New Collaboration Group** button () in the upper right corner of the page.
3. On the New Collaboration Group page, enter the **Group Name** in the appropriate text box.
4. Click **Save**. This returns you to the Manage Collaboration Groups page that lists groups that you have created in the system. When a group you create first displays here, it is empty (circled in [Figure 8.1](#)).










Manage Collaboration Groups		
2 items found, displaying all items.		
		
Collaboration Group Name	Group Members	Edit Delete
fellows 1101268119078	ResearchScientist, ResearchScientist, Xiaopeng Bian, SystemAdministrator, SystemAdministrator, caArray User, Biostatistician, Biostatistician, caArray Administrator, Leonie Mosewita, Meru Heiskanen, LabAdministrator, LabAdministrator, test user, Juh Klemm, Don Swan, Anetha Reddy, Collaborator, Collaborator	 
Laboratory JEH7	(Empty group)	 

Figure 8.1 Collaboration groups are listed on the Manage Collaboration Groups page

5. To add members to a Collaboration Group, click the **Edit** icon () corresponding to the group.
6. The group details page that opens lists any and all current members of the group. Click the **Add a New Group Member** button ().
7. In the form that opens, as appropriate, enter the registered users last name, first name, and organization (just the first letters), and click the **Filter** button () to find the user.

Note: To return all registered users, enter nothing and click the **Filter** button (not recommended due to time).

The user displays in the Member Name column.

- For each member to be added to the group, click the **Add** icon ( in the far right column of the screen.

The System automatically saves the user as a member and removes the name from the filter results.


- Click on the **Collaboration Groups** bread crumb link or the left hand menu item at the top of the page to return to the main collaboration group page .


Viewing Group Details

From the list of groups on the Collaboration Groups page, you can view all groups you have created, the first 20 members and the ability to edit or delete them. The group details lists all users in the group, along with their corresponding Institution and email address.

Editing Collaboration Group Details

To edit collaboration group details, follow these steps:

- From the Manage Collaboration Groups page, click the **Edit** icon () corresponding to the group you select.
- The page that opens lists group members, their institution and email address. On this page, you can perform the following edits:
 - Edit the Group Name. (A Group Name must be unique within the system.)
 - Add or delete group members.
 - View user details using the hypertext link corresponding to a user in the group.
 - Delete the group.
- To perform these edits, do the following:.

<i>Edit Function</i>	<i>Description</i>
Edit the group name	Enter new name in the Group Name text box.
Add a new group member	<p>Note: The new member must already have a valid caArray user account.</p> <p>Click the Add a New Group Member button at the top right of the page. In the new section of the page that opens, you can search for the group member using one or more criteria. Enter the last name, select the Role category, the Institution, and the Status to be searched. Click the Filter button.</p>
Remove a group member	On the Collaboration Group page, in the Remove column, click the icon () that corresponds to the group member.
Review group member details	Click the name of the group member. The page that opens displays contact information about the member in Read-only format.

<i>Edit Function</i>	<i>Description</i>
Delete the group	To delete the entire Collaboration Group, click the Delete button () at the bottom center of the page.

APPENDIX

A

MAGE-TAB IN CAARRAY

This appendix describes how MAGE-TAB documents are parsed, validated and imported into caArray. It also provides examples of the types of MAGE-TAB documents that are expected by caArray.

Topics in this appendix include the following:

- *caArray-Specific Handling of MAGE-TAB*
- *Best Practices and Tips* on page 105
- *Limitations of Annotation Data* on page 108
- *caArray Validation of MAGE-TAB* on page 109
- *Examples of MAGE-TAB documents* on page 115

For information about importing MAGE-TAB documents into caArray, see *Importing MAGE-TAB Data* on page 89.

caArray-Specific Handling of MAGE-TAB

Users of caArray can upload and import MAGE-TAB documents. Those topics are covered in *Chapter 6, Submitting Data to an Experiment*. See *Uploading Data Files* on page 81 and *Importing MAGE-TAB Data* on page 89.

Biomaterial Annotations in caArray

In a MAGE-TAB file, a biomaterial (Source, Sample, Extract or Labeled Extract) can be followed by any number of Characteristics columns containing annotation information about that biomaterial. Some of these Characteristics are interpreted uniquely by caArray:

- A characteristic with the OrganismPart category becomes the biomaterial's tissue site.

- A characteristic with the CellType category becomes the biomaterial's cell type.
- A characteristic with the DiseaseState category becomes the biomaterial's disease state.
- A characteristic with the Organism category becomes the biomaterial's associated organism.
- A characteristic with the ExternalSampleId category becomes the corresponding Sample's unique external identifier. You, the user, can enter this ID on the Sample details page, or as a field in a MAGE-TAB SDRF. If the user tries to enter the same external Sample ID for two different samples within the same Experiment, the System will disallow it. This identifier is guaranteed to be unique for a Sample within an Experiment, but in the future, will be unique for a Sample *across Experiments* as well. This external identifier can be used to uniquely identify a Sample across caArray as well as other applications.

Auto-Generated Missing Biomaterials in MAGE-TAB

In a MAGE-TAB SDRF being imported, if a biomaterial node is missing in the Source – Sample – Extract – Labeled Extract chain, appropriate intermediate nodes will be automatically generated to complete the chain. The number of nodes generated will depend on the left side of the graph.

Examples:

1. If the SDRF describes one source connected to three extracts, one sample will be auto-generated and inserted in the chain. On the other hand, if the SDRF describes three sources combined to generate one extract, three samples will be auto-generated and inserted into the chain.
2. If the first (leftmost) biomaterial node in the SDRF is not a source, then the chain to the left of the biomaterial node will be auto-generated by caArray. For example, say the SDRF starts with the Extract column, and there are two unique extracts in the SDRF called "liver_2900" and "liver_3200". caArray auto-generates two sources called "liver_2900" and "liver_3200", and also two samples called "liver_2900" and "liver_3200". The sources, samples and extracts are then linked in two separate chains.

Protocols Associated Intelligently

If biomaterials missing in the SDRF are auto-generated (as described in [Auto-Generated Missing Biomaterials in MAGE-TAB](#)), caArray intelligently associates protocol applications with the most appropriate auto-generated node. Note that this applies only to protocol applications in the SDRF whose association to biomaterial nodes is ambiguous. For example, if the SDRF contains a sample followed by a labeling protocol followed by a hybridization, the labeling protocol will be associated with the auto-generated extract - labeledExtract portion of the chain.

The specific rules are:

- A protocol of type "labeling" (MGED Ontology) will be associated with the extract – labeledExtract portion of the chain.

-
- A protocol of type “hybridization” (MGED Ontology) will be associated with the labeledExtract – hybridization portion of the chain.

SDRF Decides Raw versus Derived Data File

If you set the file type of a data file to raw (using the Manage Data procedures described in *Managing Data* on page 77), but the SDRF designates that same data file as derived, then the designation from the SDRF overrides the one you specified in the Manage Data interface. This also applies in the reverse case where you specify a data file to be derived in the Manage Data interface, but the SDRF specifies the data file to be raw. The designation in the SDRF is authoritative.

Image File Importing Issues in caArray

Image files cannot be validated or imported into caArray 2.2 successfully.

Best Practices and Tips

The following best practices ensure that the annotation data stored in caArray is of the highest quality.

Editing MAGE-TAB Documents

Because template MAGE-TAB IDF and SDRF files are tab-delimited, they are most easily edited in MS Excel or similar spreadsheet editing applications.

Uniquely Identifying Objects in the IDF and SDRF

An “<object> Name” header contains the unique identifier for that object in the MAGE-TAB set. E.g., Protocol Name, Experimental Factor Name and Term Source Name in the IDF; Source Name, Sample Name, Hybridization Name etc. in the SDRF.

Such a row/column should contain unique identifiers for your objects. E.g., give each of your protocols a unique identifier in the “Protocol Name” row. These unique identifiers will be used in the SDRF.

Representing Multiple Objects and Multiple Values in the IDF

Multiple Protocols, Persons, Experimental Factors and Term Sources can be defined in a single IDF file. In these cases, the different “objects” are separated by tabs.

In some cases (Person Roles, Protocol Parameters) it is possible to have multiple values within a given “object”. For example, one person may have many roles. In such cases the multiple roles should be separated by semicolons.

Protocol Parameters: If more than one parameter was used for a given protocol, the parameter names should be entered as a semicolon-delimited list. These protocol names are used in the SDRF file (as “Parameter Value [<parameter name>]” column headers) to list the values used for each protocol parameter.

Use of Controlled Vocabularies in the IDF

The terms in the IDF, listed in Table A.1 should come from controlled vocabularies:

Term	Suggested Source Ontology/Subclass
Experimental Design	MGED/ExperimentalDesignType
Experimental Factor Type	MGED/ExperimentalFactor Category Note: If in the MGED Ontology under the ExperimentalFactorCategory, Class and Individual types appear to be the same, the Class is used as a column header, whereas the Individual is used as an instance of a class. For example, Class "DiseaseState" and Individual "disease_state". The column header is Characteristics [Disease State]; the individual is Experimental Factor Type "disease-state".
Person Roles	MGED/Roles, multiple values as semicolon delimited list. Examples: submitter, investigator
Quality Control Type	MGED/QualityControlDescriptionType
Replicate type	MGED/ReplicateDescriptionType
Normalization Type	MGED/NormalizationDescriptionType
Protocol type	MGED/ProtocolType

Table A.1 Controlled vocabularies in the IDF

Note: If you cannot find your term in any ontology, you can enter the term you want in the Term Source REF column.

Term Source REF in an SDRF

To insure that annotation data stored in caArray is of the highest quality, it is important that Term Source REFs be included for everything possible. For example, every Characteristics column should be followed by a Term Source REF column detailing the ontology the term came from. As far as possible, the MGED Ontology and the NCI Thesaurus should be used as Term Source REFs.

Each entry in the Term Source REF column should have a corresponding entry in the IDF.

Note: If the term source is unknown, enter *caArray* in the Term Source Ref column.

Use of Controlled Vocabularies in the SDRF

The columns in the SDRF should contain values from controlled vocabularies, described in Table A.2. .

Term	Suggested Source Ontology/Subclass
Characteristics []	This column contains terms describing each material according to the category indicated in the column header. E.g., a "Characteristics [OrganismPart]" column would contain individual OrganismPart terms from the NCI Thesaurus or other ontology sources. Alternatively, the Characteristic could be a measurement, indicated using a Unit [] column following it.
Material Type	The values should typically come from the MGED Ontology, from the "MaterialType" class.
Label	The values should typically come from the MGED Ontology, from the "LabelCompound" class.
Factor Value []	Factor value should either be an ontology term, such as "Brain" or "Breast Cancer" from NCI Thesaurus, or a number followed by a "Unit" column such as "hours", "mg", etc.
Unit []	The category and values should typically come from the MGED Ontology, from any of the subclasses of "Unit".
Term Source REF	Each entry in this column should have a corresponding entry in the IDF. See <i>Term Source REF in an SDRF</i> on page 106.

Table A.2 Columns in an SDRF that contain terms from controlled vocabularies

Biomaterials Column Order in an SDRF

In an SDRF file, biomaterials columns must follow the order described in Table A.3. Note that all column types, such as Comment, are not mandatory in the SDRF.

Initial Biomaterial Name Column	Column Order after the First Initial Biomaterial Name Column
Source Name	Provider, Material Type, Characteristics [], and Protocol REF. Description and Comment are optional; most users do not include them.
Sample Name	Material Type and Characteristics. Description and Comment are optional; most users do not include them.
Extract Name	Material Type and Characteristics. Description and Comment are optional; most users do not include them.
Labeled Extract Name	Label, Material Type and Characteristics. Description and Comment are optional; most users do not include them.

Table A.3 Column order of biomaterial in an SDRF

References from the SDRF to the IDF

Each of the SDRF columns listed in Table A.4 contain references to values in the IDF file.

SDRF Column	Description of Reference to IDF File
Protocol REF	Contains references to Protocol Name values defined in the IDF. Multiple protocols can be chained together by placing together multiple Protocol REF columns in the appropriate order. For example, the user's SDRF has a Sample followed by an extraction protocol followed by a labeling protocol followed by a labeled Extract column. caArray auto-generates the missing Extract and associate the protocols to the appropriate portion of the chain.
Parameter Value []	One or more columns that follow a Protocol REF column in the SDRF; refer to the Protocol Parameters defined in the IDF for that protocol.
Factor Value []	One or more columns in the SDRF that represent the experimental factor values for a hybridization, and reference Experimental Factor Names defined in the IDF.
Term Source REF	Each Term Source REF column in the SDRF contains references to a Term Source Name defined in the IDF. See <i>Term Source REF in an SDRF</i> on page 106.

Table A.4 References in an SDRF to an IDF

Limitations of Annotation Data

The following annotations that can be imported into the caArray repository via MAGE-TAB are not visible on the caArray application user interface:

- Provider (of a Source)
- Date of Experiment
- Public Release Date
- Person Mid Initials, Fax, Address and Affiliation
- Publication DOI (Digital Object Identifier)
- Protocol Parameters
- Unit (of a Characteristic, Parameter Value or Factor Value)
- Performer (of a Protocol)
- Protocol Date
- Factor Value

Some annotations that can be imported into the repository via MAGE-TAB are visible but not editable in the caArray application user interface:

Characteristics are all read-only except for the OrganismPart, CellType, DiseaseState and Organism categories for a Source.

caArray Validation of MAGE-TAB

MAGE-TAB IDF Fields Recognized


In caArray, the following headers are recognized in an IDF file. Any other header results in a validation error.

1. Investigation Title
2. Experiment Description
3. Date of Experiment
4. Public Release Date
5. Experimental Design
6. Experimental Design Term Source REF
7. Experimental Factor Name
8. Experimental Factor Type
9. Experimental Factor Term Source REF / Experimental Factor Type Term Source REF
10. Person Last Name / Person Last Names
11. Person First Name / Person First Names
12. Person Mid Initials / Person Mid Initial
13. Person Email
14. Person Phone
15. Person Fax
16. Person Address
17. Person Affiliation
18. Person Roles
19. Person Roles Term Source REF
20. Quality Control Type / Quality Control Types
21. Quality Control Term Source REF / Quality Control Types Term Source REF
22. Replicate Type
23. Replicate Term Source REF / Replicate Type Term Source REF
24. Normalization Type
25. Normalization Term Source REF
26. PubMed ID

27. Publication DOI
28. Publication Author List
29. Publication Title
30. Publication Status
31. Publication Status Term Source REF
32. Protocol Name
33. Protocol Type
34. Protocol Description
35. Protocol Parameters
36. Protocol Hardware
37. Protocol Software
38. Protocol Contact
39. Protocol Term Source REF
40. SDRF File / SDRF Files
41. Term Source Name
42. Term Source File
43. Term Source Version
44. Comment – Parsing of this element is deferred.

MAGE-TAB IDF Validation Rules

Validation **Errors** will occur in caArray if any of the following rules are broken in the IDF files.

1. There must be only one IDF in the file set being validated.
2. Illumina Data CSV and Genepix GPR files contain hybridization/sample names that are implicit. The SDRF validation will fail if it does not contain these hybridization/sample names. 
3. Integrity of references:
 - a. An Experimental Factor Type must not refer to a non-existent Experimental Factor Name.
 - b. If the header SDRF File / SDRF Files exists, then it must have at least one value.
 - c. Any referenced SDRF file must exist in the file set being validated.
4. Formats:

-
- a. Date of Experiment and Public Release Date values must be in the format YYYY-MM-DD.

Validation **Warnings** will occur in caArray if any of the following rules are broken in the IDF files.

5. If one of the following headers exists, then it must have exactly one value:
 - a. Investigation Title
 - b. Experiment Description
 - c. Date of Experiment
 - d. Public Release Date
6. If one of the following headers exists, then it must have at least one value:
 - a. Experimental Design
 - b. Experimental Factor Name
 - c. Person Email
 - d. Person Phone
 - e. PubMed ID
 - f. Publication Title
 - g. Protocol Name
7. Integrity of references:
 - a. For every "<element> Term Source REF", a corresponding Term Source Name must be defined in the IDF. Otherwise a default "caArray" term source is assumed.
 - b. For every "<element> Term Source REF", a corresponding "<element>" must be defined in the IDF. E.g., for every Experimental Factor Term Source REF, a corresponding Experimental Factor Name must be defined in the IDF.
 - c. For every Person Roles Term Source REF, a corresponding Person Name must be defined in the IDF.

MAGE-TAB SDRF Fields Recognized

The following headers are recognized by caArray in an SDRF file. Any other header will result in a validation error.

1. Source Name
2. Sample Name
3. Extract Name
4. Labeled Extract Name

5. Hybridization Name
6. Scan Name
7. Normalization Name
8. Provider
9. Protocol REF
10. Characteristics
11. Material Type
12. Parameter Value
13. Term Source REF
14. Unit
15. Label
16. Array Design File (Not supported. If this column exists, it will result in an error. Use Array Design REF instead.)
17. Array Design REF
18. Array Data File
19. Derived Array Data File
20. Array Data Matrix File
21. Derived Array Data Matrix File
22. Image File
23. Factor Value
24. Performer
25. Protocol Date
26. Description
27. Comment – Parsing of this element is deferred.

MAGE-TAB SDRF Validation Rules

Validation **Errors** will occur in caArray if any of the following rules are broken in the SDRF files.

1. All three of the following columns must be present:
 - a. Biomaterial (Source Name, Sample Name, Extract Name or Labeled Extract Name)
 - b. Hybridization Name

-
- c. File (Array Data File, Derived Array Data File, Array Data Matrix File or Derived Array Data Matrix File)
 2. If one of the following columns exists, then it must not be blank:
 - a. Source Name
 - b. Sample Name
 - c. Extract Name
 - d. Labeled Extract Name
 - e. Hybridization Name
 - f. Array Data File
 - g. Array Data Matrix File
 - h. Derived Array Data File
 - i. Derived Array Data Matrix File
 3. Only one instance of the following columns can exist:
 - a. Source Name
 - b. Sample Name
 - c. Extract Name (Supporting multiple Extract Name columns is deferred.)
 - d. Labeled Extract Name
 - e. Hybridization Name
 - f. Scan Name
 - g. Normalization Name
 - h. Array Data File
 - i. Array Data Matrix File
 4. An Array Design Name column is unsupported. Use an Array Design REF instead.
 5. Column ordering:
 - a. The following columns, although not mandatory, where present, must occur in the following order, starting left and proceeding to the right (not necessarily consecutively): Source Name, Sample Name, Extract Name, Labeled Extract Name, Hybridization Name, Scan Name, Raw data, Normalization Name, Derived data. ("Raw data" refers to any number of Array Data File and/or Array Data Matrix File columns. "Derived data" refers to any number of Derived Array Data File and/or Derived Array Data Matrix File columns.)
 - b. A Provider must occur to the right of a Source Name. It must not occur to the right of any of the other biomaterial columns.

- c. A Material Type must occur to the right of a biomaterial column.
 - d. A Parameter Value column must occur to the right of a Protocol REF column.
 - e. A Unit must immediately follow a Characteristic, Parameter Value or Factor Value column.
 - f. A Description must be immediately preceded by a biomaterial column.
 - g. A Label must occur to the right of a Labeled Extract Name.
 - h. A Factor Value must occur to the right of a Hybridization Name.
 - i. An Image File must occur to the right of a Hybridization Name.
6. Integrity of references:
- a. For a Term Source REF, the corresponding Term Source Name must be defined in the IDF.
 - b. For a Protocol REF, the corresponding Protocol Name must be defined in the IDF.
 - c. An Array Design REF value must refer to the LSID of an array design that has already been imported into the System. The System uses LSIDs in the format "URN:LSID:authority:namespace:object". E.g., URN:LSID:Affymetrix.com:PhysicalArrayDesign:HG-Focus.
 - d. A referenced data file (Array Data File, Array Data Matrix File, Derived Array Data File or Derived Array Data Matrix File) must be in the set being imported.
 - e. For a Factor Value, the corresponding Factor Name must be defined in the IDF.

Validation **Warnings** will occur if any of the following rules are broken in the SDRF files.

- 1. Integrity of references:
 - a. For a Protocol REF, the corresponding Protocol Name must be defined in the IDF.
- 2. Column ordering:
 - a. If a Scan Name does not occur between a Hybridization Name and a Raw data column (Array Data File or Array Data Matrix File), it will be ignored.
 - b. If a Normalization Name does not occur between a Raw data column and a Derived data column (Derived Array Data File or Derived Array Data Matrix File), it will be ignored.
 - c. A Performer must occur to the right of a Protocol REF, otherwise it will be ignored.
- 3. Formats:
 - a. Protocol Date values must be in the format YYYY-MM-DD.

Examples of MAGE-TAB documents

The IDF-SDRF examples provided by NCICB are in a single zip file called `magetab_samples.zip` in NCICB GForge, located under **Docs > User Community** at this site: <https://gforge.nci.nih.gov/docman/view.php/305/13249/>.

APPENDIX B

CAARRAY REFERENCES

- **NCICB Production Site:** <https://array.nci.nih.gov>
- **Product Summary Site:** <https://cabig.nci.nih.gov/tools/caArray> - the summary of caArray capabilities and direction
- **Public Information Site:** <http://caarray.nci.nih.gov/> – a public web site that allows anyone to download the latest version, access documentation, launch the portal and visit sites that provide analysis of the data contained in caArray.
- **caArray Work Group Site:** <https://cabig.nci.nih.gov/workspaces/ICR/caArray-wg/> – this public web site provides access to the schedule, monthly meeting notes and links to the listserv for the stakeholder community
- **Microarray Gene Expression Data Society -** <http://mged.org/> The providers and curators of microarray standards, software and models.
- **MAGE-TAB Specification:** <http://www.mged.org/mage-tab/spec1.0.html>
- **Tab2MAGE ArrayExpress package:** <http://tab2mage.sourceforge.net/> - MAGE-TAB examples
- **MGED Ontology:** <http://mged.sourceforge.net/ontologies/MGEDontology.php>
- **NCI Thesaurus:** <http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do>
- **NCBI Taxonomy (ncbitax):** <http://www.ncbi.nlm.nih.gov/Taxonomy/>

INDEX

A

account, requesting new user 10

adding

array designs 65

group 99

publication 54

vocabulary term 51, 52, 73

annotations

adding vocabulary 51

biomaterials 103

Annotations tab 34

API, extracting data 95

Application Support i, 16

array designs

adding 65

deleting 68

editing 68

managing 64

viewing 64

B

biological source materials

description 37

relationship between 38

biomaterials

annotations 103

autogeneration when MAGE-TAB data

imported 90, 104

columns in SDRF 107

missing in SDRF 104

browsing

after login 9

before login 8

caArray 8, 19

C

caArray

About caArray sidebar links 16

browsing 8, 19

common uses 16

compatible browsers 7

Data Portal page 10

interface description 14

launching 7

login 8

My Experiment Workspace 17

objectives 5

overview 5

relationship to caBIG 6

requesting user account 10

sample search results 27

searching the repository 22

search results 26

task tabs 17

user interface footer 16

Welcome page 15

Welcome to caArray sidebar links 15

caArray User's Guide

introduction 1

organization 1

text conventions 2

caBIG, relationship to caArray 6

collaboration group

creating 99

editing 100

viewing details 100

collaboration groups

assigning permissions selectively 59

managing 99

setting experiment visibility 60

Contact tab 33

controlled vocabularies, use in IDF 106

controlled vocabularies, use in SDRF 107

copying

biomaterials 50

hybridizations 50

creating

collaboration group 99

experiment 30

group 99

protocol 42, 44, 45, 47, 70, 71

user account 98

curation tasks 63

D

data

- uploading, file sizes 81

Data portal page 10

Data tab

- deleting a file 83
- description 53
- downloading data files 93
- importing files 86
- Manage Data subtab 77
- Supplemental files 91
- uploading data 81
- validating data 83

data tab

- filtering file display 80

deleting

- array design 68
- biomaterials 50
- data file 83
- experiment 56
- hybridizations 50

derived file vs. raw file 105

description 91

downloading

- annotation files, source 51
- data files 93
- data set size 94
- MAGE-TAB data 91, 93, 95

draft status, Experiment visibility in 58

E

editing

- array designs 68
- collaboration group 100
- MAGE-TAB documents 105
- protocol 71
- vocabulary term 74

editing protocol type 72

experiment

- adding vocabulary 51
- Annotation tab 34
- Contact tab 33
- creating 30
- Data tab 79
- deleting 56
- draft status 58
- generating MAGE-TAB files 94
- in progress 58
- managing data submission 77
- Overview tab 31
- Publications tab 54
- searching for 23
- status 55

- submitting 55

- updating 56

- uploading data to 81

- visibility 57, 60

Experimental Factors tab 37

experiments, overview 29

extract

- adding new 43

- description 38

extracting

- caArray files 93

- data programmatically 95

Extracts tab 43

F

files, filtering on Data tab 80

file types, importable 78

filtering

- by file status 94

- by file type 94

G

Global Quick Links 16

grid availability 95

group

- creating 99

- viewing details page 100

H

help in caArray 13

hybridization

- adding new 46

- viewing 48

Hybridizations tab 46

I

IDF

- headers recognized in caArray 109

- representing multiple objects in 105

- representing multiple values in 105

- uniquely identifying objects in 105

- use of controlled vocabularies in 106

- validation rules in caArray 110

image

- cannot import files 79

- issues with importing 105

importable data file types 78

Imported Data tab 88

importing

- data files 86

- MAGE-TAB data 89

- MAGE-TAB SDRF 90, 104

in progress status, visibility in 58

L

labeled extract

adding new 44

description 38

Labeled Extracts tab 44

launching caArray 7

login 10

M

MAGE-TAB

data files 89

downloading data 95

editing tips 105

generating IDF, SDRF files from caArray
experiment 94

IDF fields recognized in caArray 109

IDF validation rules in caArray 110

import

with existing experiment components 90

importing data 89

SDRF fields recognized in caArray 111

SDRF validation rules in caArray 112

validating in caArray 109

managing

array designs 64

controlled vocabulary terms 40

data 77

user accounts 98

user groups 99

vocabulary terms 72

My Experiment Workspace 17

N

NCICB Application Support i, 16

O

online help

icons 13

using 13

overview, chapters in guide 1

Overview tab 31

P

permissions

assigning by sample only 59

configuring 57

protocol

creating 42, 44, 45, 47, 70, 71

editing 71

protocols

description 68

in SDRF 104

viewing 68

protocol types

editing 72

viewing 70

publication, adding 54

Publications tab 54

R

raw file vs derived file 105

registering new user 10

S

sample

adding new 40

description 38

samples

searching 24

Samples tab 40

SDRF

associating protocols in 104

biomaterials columns in 107

fields recognized in caArray 111

import 104

uniquely identifying objects in 105

use of controlled vocabularies in 107

validation rules in caArray 112

searching

caArray repository 22

experiments 23

from caArray login screen 8

samples 24

search results

experiments 26

samples 27

source

adding new 39

description 38

downloading annotation files 51

source materials

annotations for 38

characteristics 38

definition 38

Sources tab 38, 39

status of experiment 55

submitting experiment 55

Supplemental files

file types defined as 83

supplemental files 91

T

TCGA policy 58

Technical Support [i, 16](#)
text conventions in user guide [2](#)

U

uploading data
 file size issues [81](#)
 file types [81](#)
 to experiment [81](#)
user's manual conventions [2](#)
user account, login [10](#)
user account, new [98](#)
user groups
 editing [100](#)
 managing [99](#)

V

validating
 data file [83](#)
 MAGE-TAB in caArray [109](#)
validating data
 steps in [84](#)
validation errors [85](#)
viewing
 array designs [64](#)
 collaboration group details [100](#)
 hybridization [48](#)
 protocols [68](#)
 protocol types [70](#)
 vocabulary term [72](#)
visibility
 assigning selectively [59](#)
 collaboration groups, setting [60](#)
 Experiment in draft status [58](#)
 experiment in progress status [58](#)
vocabulary term
 adding [51, 52, 73](#)
 editing [74](#)
 viewing [72](#)
vocabulary terms, managing [72](#)