caArray 2.4

User's Guide

NATIONAL® CANCER INSTITUTE

Center for Biomedical Informatics
and Information Technology

July 27, 2010

CREDITS AND RESOURCES

caArray Development and Management Teams			
Development	***Quality Assurance***	***Documentation***	***Project and Product Management***
Rashmi Srinivasa[2]	Xiaopeng Bian[1]	Rashmi Srinivasa[2]	Maki Duncan[2]
Dan Kokotov[2]	Quy Phung[6]	Dan Kokotov[2]	Rashmi Srinivasa[2]
Ashenafi Gax Ayalew[2]	Matthew Tiller	Doug Harley[2]	Anand Basu[1]
Doug Harley[2]		Jill Hadfield[1]	Juli Klemm[1]
Systems and Application Support		***Training***	
Yeon Choi[3]		Don Swan[3]	
Andrea Johnson[3]			
Cuong Nguyen[3]			
Jacob Mensah[3]			
[1] NCI Center for Biomedical Informatics and Information Technology (CBIIT)		[2] 5AM Solutions	[3] Terrapin Systems
[4]NARTech	[5]Northern Taiga Ventures, Inc.	[6]Enterprise Solutions and Consulting	

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

TABLE OF CONTENTS

USING THE CAARRAY USER'S GUIDE

This chapter introduces you to the *caArray 2.4 User's Guide* and suggests ways you can maximize its use.

Topics in this chapter include:

Introduction to the caArray User's Guide

The *caArray 2.4 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 2.4 User's Guide* contains the following chapters and appendices:

***Using the caArray User's Guide*—**This chapter introduces you to the *caArray 2.4 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 **Save** button () to save the file.
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.
Italics	Used to reference other documents, sections, figures, and tables.	*caCORE Software Development Kit 1.0 Programmer's Guide*

Table 2.1 caArray Guide Text Conventions

Convention	*Description*	*Example*
Italic boldface monospace type	Text that you type	In the New Subset text box, enter ***Proprietary Proteins.***
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

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

caArray 2.3 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 2.4 at your site, see your local administrator and/or refer to the *caArray Local Installation Guide* 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*).

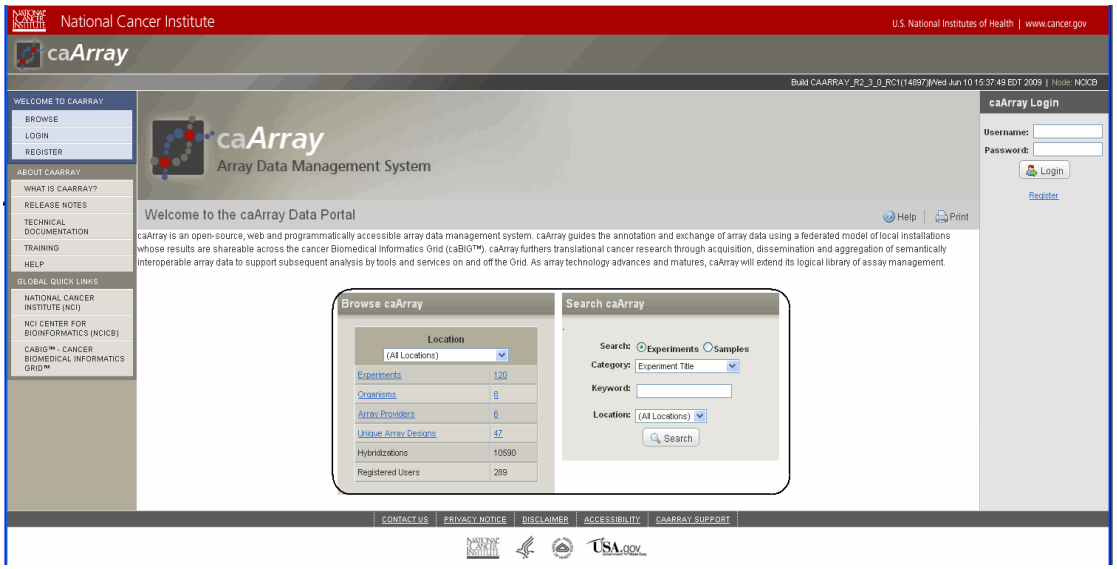


Figure 2.1 Prior to login, you can use these dialogs to browse or search the caArray database

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 55.

- 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 *Registering as a New caArray User* 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*).

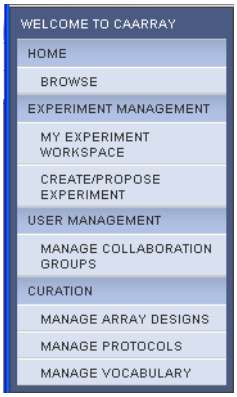


Figure 2.2 Browse options on left sidebar

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

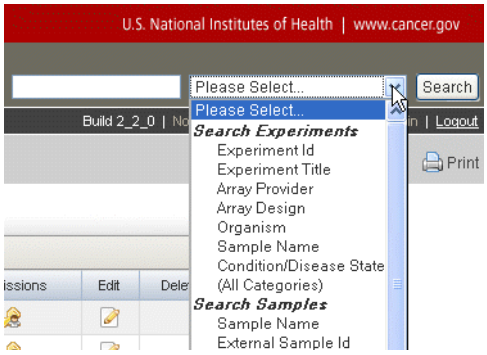


Figure 2.3 Search text box displays in every browser page

For more information about launching a search, see *Chapter 3 Navigating and Searching caArray*.

caArray User Accounts and Login

From the caArray Data Portal Home/Login page you can you can you can log in, if you have an existing account, or you can request a new user account.

To log in, see *Logging into caArray* on page 10.

To register for a new user account, see *Registering as a New caArray User* on page 10.

Logging into caArray

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*):

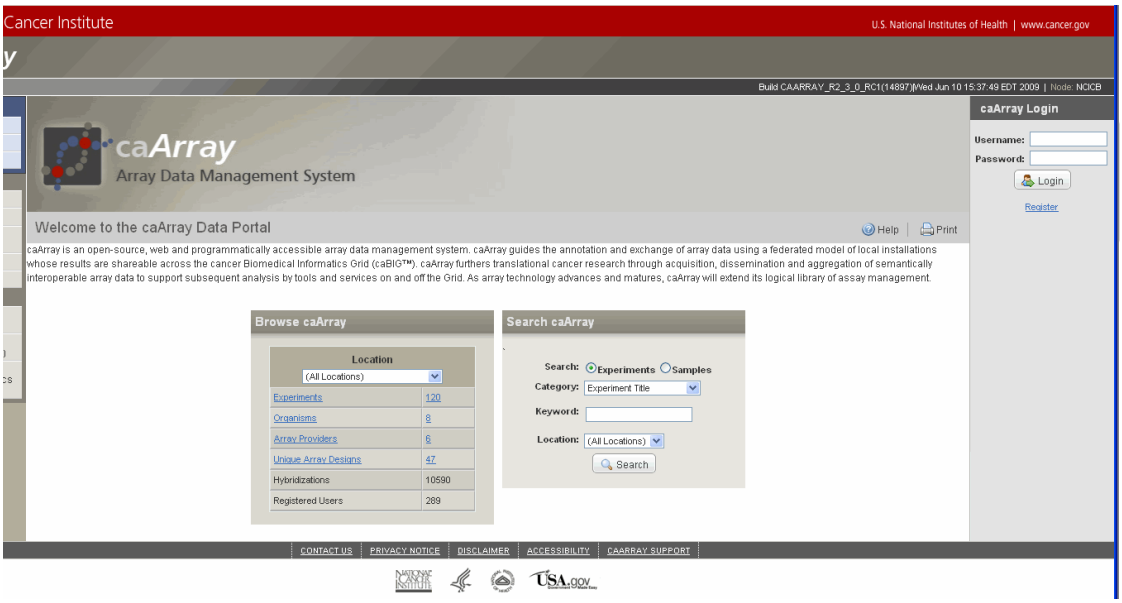


Figure 2.4 caArray login page

To 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.

Registering as a New caArray User

To request a caArray user account, you must register as a new user, completing the following steps:

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*).

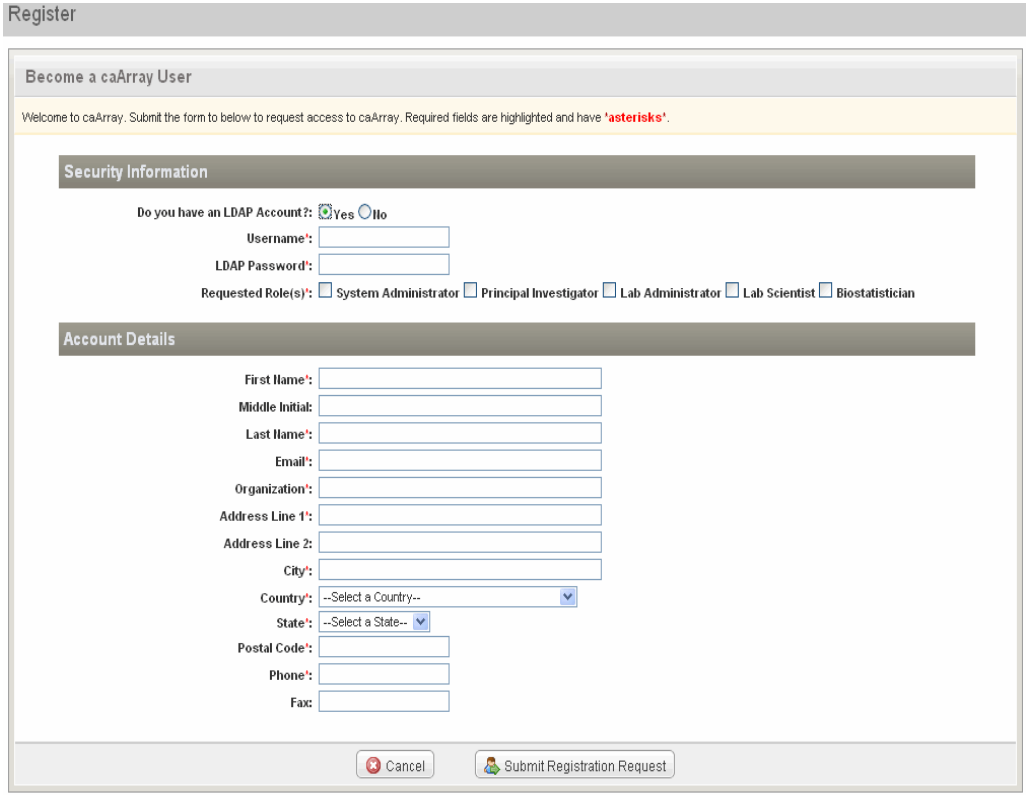


Figure 2.5 New user account registration form

3. In the Become a caArray User form, enter the appropriate information[1].

   - **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*

     –LDAP 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.

   This opens a Registration Request confirmation page.

At this point, an email containing all of the information you specified in the new user request form is sent to the caArray system administrator and an account request confirmation email is also sent to you, the prospective user, at your specified email address. In response, the caArray system administrator uses UPT to create your user account and assign the requested roles (in predefined groups like Principal Investigator). When your account is created, the system administrator sends you an email to alert you, after which you can login to caArray.

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.

For information about administering user accounts, see *Administering caArray User Accounts Using UPT* on page 97

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 breadwinner 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 2.4 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 2.4 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 *Searching the caArray Repository* 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

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.

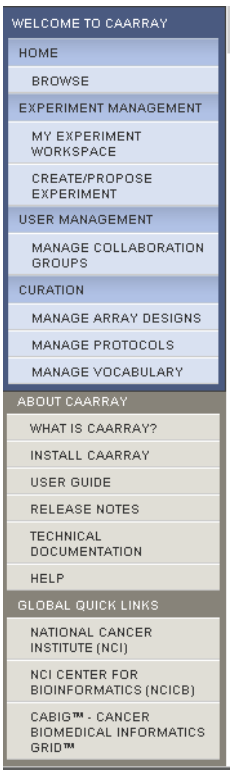


Figure 2.6 The caArray Welcome page navigation or left sidebar menu

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 *Administering caArray User Accounts Using UPT* on page 97.
  - **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 options relating to the experiments with which you are associated.

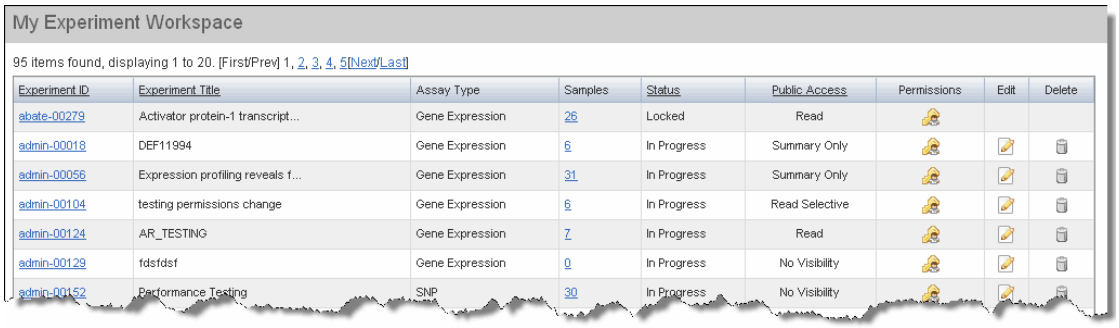


Figure 2.7 caArray My Experiment Workspace

The My Experiment workspace lists experiments with their status (“in Progress” or “Locked”). The Locked experiemnts do not have an Edit or Delete button next to them. For more information, see *Experiment Visibility* on page 55.

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: **In Progress** or **Locked**
- **Permissions** – Click the icon to assign or modify the experiment permissions. See *Experiment Visibility* on page 55.
- **Edit** – Click the icon to edit experiments with the appropriate permissions. See *Editing an Experiment* on page 54.

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

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 *Searching the caArray Repository* on page 22.

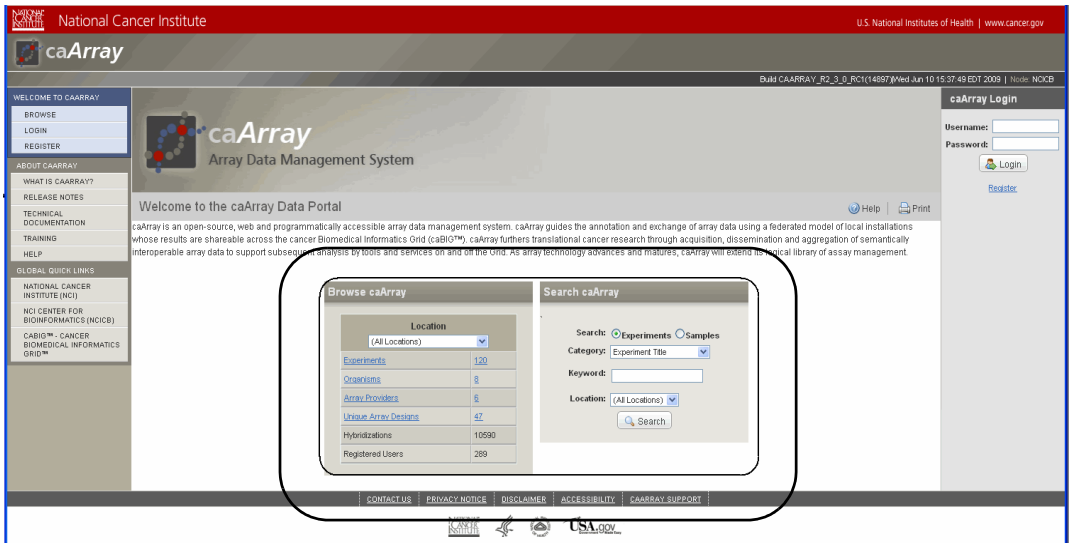


Figure 3.1 Prior to login, you can use these dialog boxes to browse or launch a search of the caArray database

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. *About File Types in caArray* on page 76

   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	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. 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. **Note:** Only Affymetrix, Illumina and GenePix formats are fully supported with validation and parsers in caArray 2.2. For more information, see *About File Types in caArray* on page 76 and *Managing Data* on page 75.

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

Table 3.1 Browse dialog box categories (Continued)

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*
Experiment ID	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.
Experiment Title	The experiment title defined manually, naming and/or briefly describing the experiment
Assay Type	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.
Organism	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.
Updated	The date of the most recent update of the experiment

Table 3.2 Experiment metadata categories (Continued)

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:*Browsing and Searching Before Login* on page 8*Browsing and Searching Before Login* on page 8

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*).

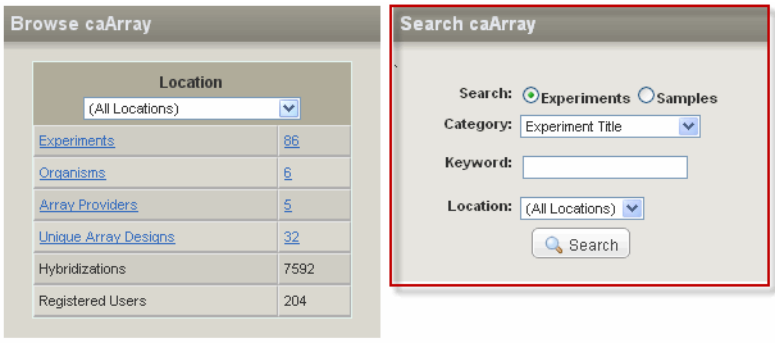


Figure 3.2 Search dialog box

- OR -

2. *After you log in*, locate the Search area of the page*Browsing and Searching Before Login* on page 8, in the upper right-hand corner.

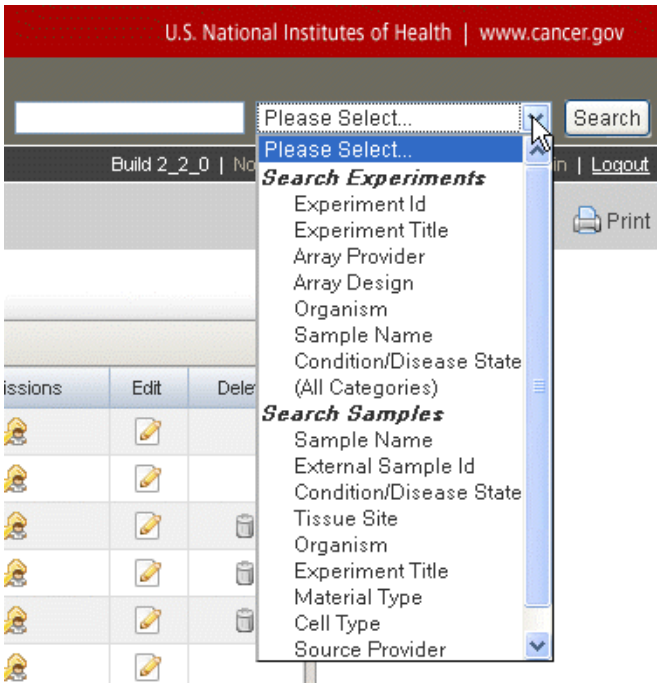


Figure 3.3 Section of every caArray page for launching a search

3. Define the search criteria by using the search options described *Table 3.3*.:

Search Option	***Description***
Keyword	In the text box, enter one or more words, separated by spaces. *Example*: ***breast cancer*** **Note:** If you leave the text box empty, caArray prompts you to enter at least two characters, unless you choose **Organism** or **Other Characteristics** fields in which cases the Keyword field may be left blank. If the Keyword field is left empty, caArray returns all samples/sources that have the chosen characteristic. 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.

Table 3.3 Search criteria options

Search Option	*Description*
Search Experiments – Category	This option is used in conjunction with keyword(s) you enter in the Keyword field. Select one of the following Experiment property categories: • **Experiment ID** • **Experiment Title** • **Array Provider** • **Array Design** • **Organism** • **Sample Name** • **Condition/Disease State** • **All Categories** 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.
Search Samples – Category	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 require keywords. If the keyword field is left empty, caArray returns all samples/ sources that have the chosen characteristic. Select one of the Samples property attributes. All of the sample search options are predefined, except for **Other Characteristics**, described below. • **Sample Name** • **Sample External ID** • **Condition/Disease State** • **Tissue Site** • **Organism** • **Experiment Title** • **Material Type** • **Cell Type** • **Source Provider** • **Other Characteristics*** • **All Categories** 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.

Table 3.3 Search criteria options (Continued)

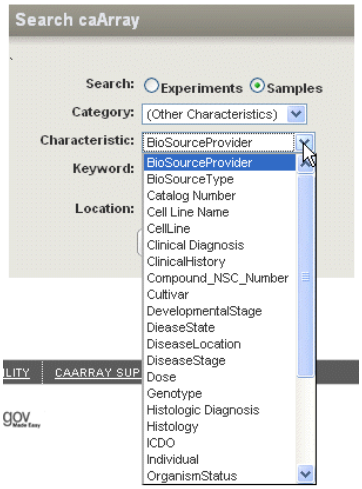
Search Option	*Description*
Search Samples – Category (cont'd)	*When you select **Other Characteristics**, an additional set of attrributes displays in the drop-down menu (*Figure 3.4*). These are arbitrary source or sample characteristics imported through MAGE-TAB SDRF files. **Note:** The **Other Characteristics** option is available only when you launch a search from the caArray Main Page.*Browsing and Searching Before Login* on page 8.  *Figure 3.4 Source or Sample characteristics imported as part of MAGE-TAB SDRF file*
Location	The list displays only the current caArray instance you are using, either your local institution or CBIIT.

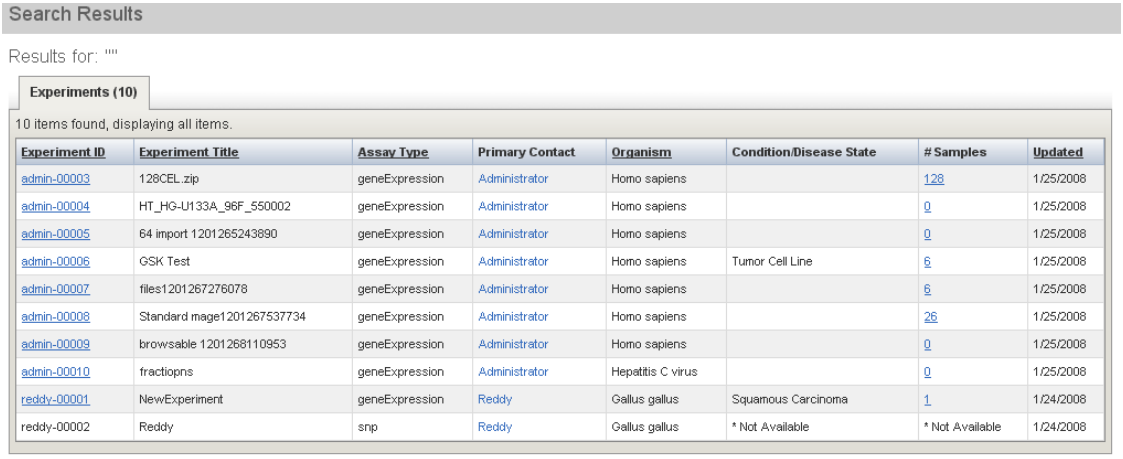
Table 3.3 Search criteria options (Continued)

4. 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	fractiopns	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.

Search Results Properties	***Search Results Fields Descriptions***
Experiment ID	The auto-generated identification assigned by caArray. Click the hypertext link to open the corresponding experiment tabs which contain all current experiment information.
Experiment Title	The experiment title defined manually, naming and/or briefly describing the experiment
Assay Type	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. Click the hypertext link or the envelope icon ( ) to open an email form where you can draft an email to this contact, if named.
Organism	The organism that is the source of the sample biomaterials used in the experiment

Table 3.4 Experiment metadata categories

Search Results Properties	Search Results Fields Descriptions
Disease State	The disease state of the source materials used in the experiment
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.
Updated	The date of the most recent update of the Experiment draft

Table 3.4 Experiment metadata categories (Continued)

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 *Editing an Experiment* on page 54.
- For information about contacting the experiment POC, see *Primary Contact* in *Table 3.4*.
- For information about extracting data from an experiment, see *Downloading Data from caArray* on page 91.

Tip: 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.

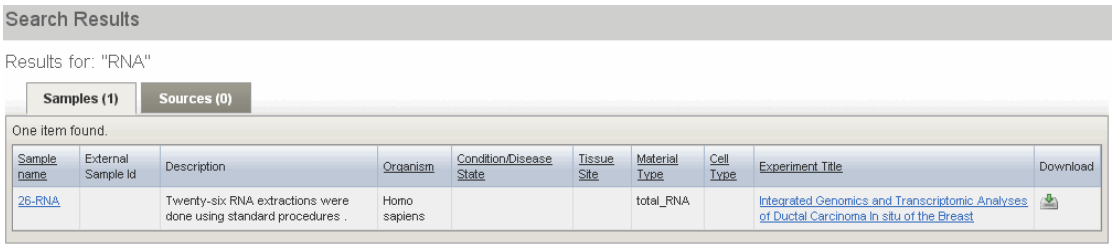


Figure 3.6 Sample/Sources Results tab

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 *Setting Selective Permissions* on page 56.

Search results are listed in table format, with columns displaying properties for each sample or source; fields are described in *Table 3.5*. 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.

Search Results Properties	***Search Results Fields Descriptions***
Sample/Source Name	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 *Sources Tab* on page 37 or *Samples Tab* on page 39.
Sample External ID	Identifier given to a sample in addition to its name.
Description	Description of sample or source
Organism	The organism that is the source of the sample biomaterials used in an experiment.
Condition/Disease State	The disease state of the source materials used in an experiment
Tissue Site	The site from which the source material was obtained.
Material Type	The type of source material being used for an experiment.
Cell Type	The category of cells used for the source material.
Provider	The organization that provided a source from which the sample was derived.
Experiment Title	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.
Download	Click the **Download** icon ( to download the data files associated with the sample or source. For information about extracting data from an experiment, see *Downloading Data from caArray* on page 91.

Table 3.5 Sample or Source metadata categories

Tip: 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

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 (Continued)

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 79.

With the appropriate permissions, you can create ("propose") an experiment, save the draft, edit it, and finally publish the experiment with its corresponding annotations.

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 Characteristics[2] as described in the *Table 4.2*.

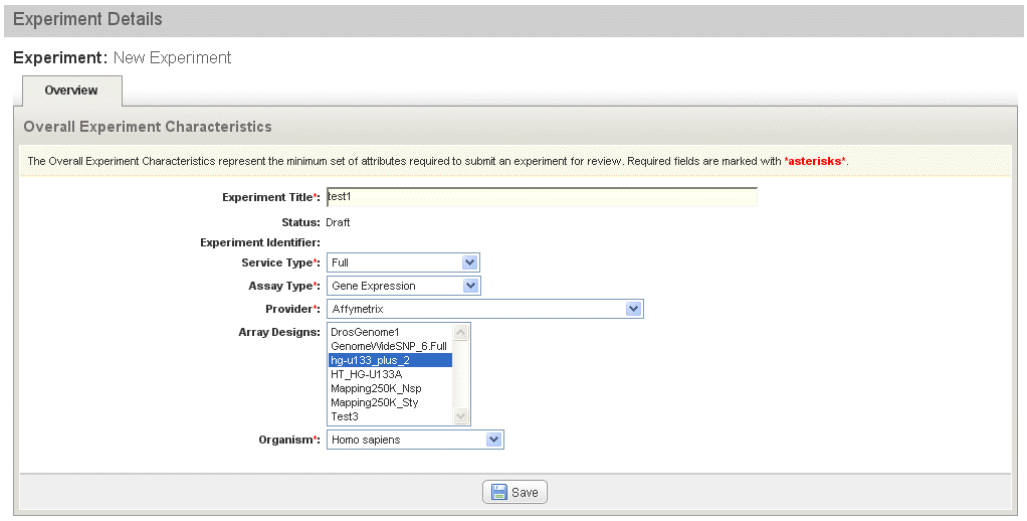


Figure 4.2 : Overview tab for an Experiment

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	The default status of an experiment is “In Progress”.

Table 4.2 Fields for Overall Experiment Characteristics

2. Fields with a red asterisk * are required.

Overview Tab Fields	*Description*
Experiment Identifier	This project identifier is autogenerated by caArray upon the initial save of the experiment. The experiment identifier is not editable. • IDs generated prior to caArray v.2.4 were composed of the PIs last name followed by a 5 character number. *Example*: jdoe-90765. After the experiment has been saved or submitted, the experiment ID displays as a hypertext link that opens the experiment. • IDs generated with v.2.4 are composed of the string "EXP-" followed by a number (no limit to digits). Already existing experiment public identifiers remain unchanged.
Assay Type*	Select from the drop-down menu the appropriate assay type. Options are the following: • **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*	Select from the drop-down menu the provider of the array. **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 *Managing Data* on page 75. Once selected, caArray automatically loads a corresponding list of array designs (next field).
Array Designs	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 *Managing Array Designs* on page 62.
Organism*	Select from the drop-down menu the organism that is the source of the sample biomaterial used in the experiment.

Table 4.2 Fields for Overall Experiment Characteristics (Continued)

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*.

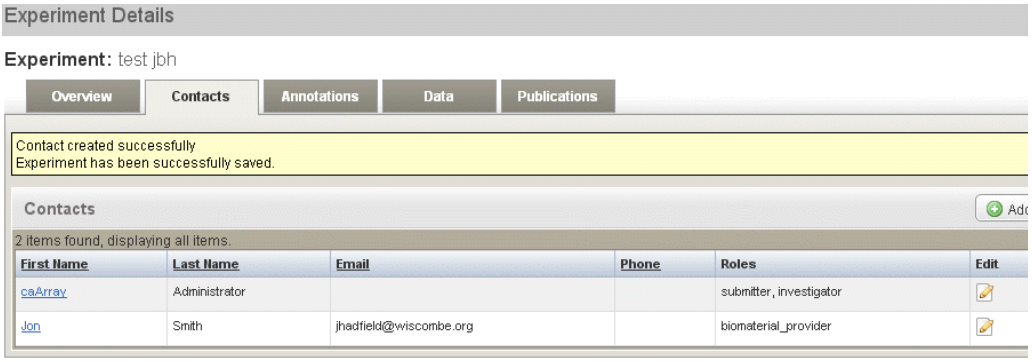


Figure 4.3 A cropped version of the Contacts tab

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
Roles*	Select the role performed by the contact you are adding and click the adjoining plus icon ( ) to move it into the **Selected Roles** panel. 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.

Table 4.3 Contact fields

2. Click **Save**.
3. 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.
4. 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.

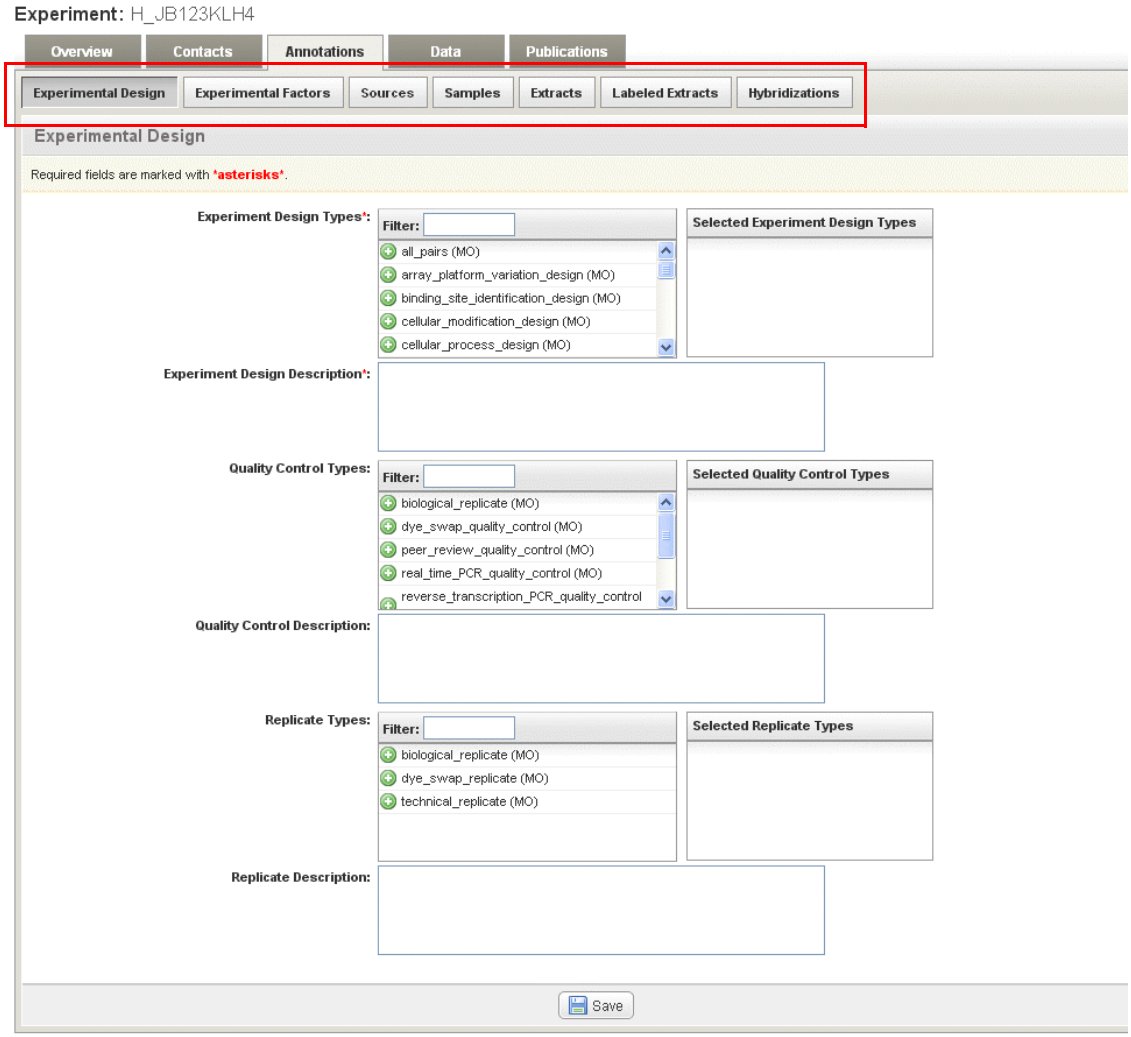


Figure 4.4 Upon opening, the Annotations tab displays the Experimental Design subtab and 6 other subtabs for entering experiment annotation data.

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*	If the appropriate Experimental Design Type displays in the left panel, click the adjoining Plus icon ( ) to move it into the **Selected Experimental Design Types** panel. 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.
Experimental Design Description	Enter a description for the experimental design used for the experiment.
Quality Control Types	Select the QC type in the displayed list. If the appropriate QC Type displays in the left panel, click the adjoining Plus icon ( ) to move it into the **Selected QC Types** panel. 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.
Quality Control Description	Enter a description for the quality control used for the experiment.
Replicate Types	Select one or more replicate types from the displayed list. Replicates can be either technical (arrays) or biological (laboratory animals or samples, etc.) If the appropriate Replicate Type displays in the left panel, click the adjoining Plus icon ( ) to move it into the **Selected Replicate Types** panel. 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.
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 87.

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].

Experimental Factors Fields	***Description***
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.

4. Fields with a red asterisk * are required.

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 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.

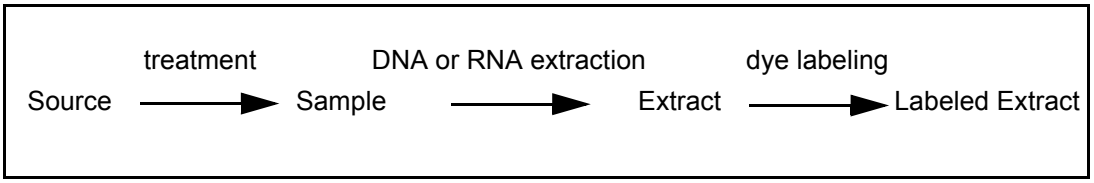


Figure 4.5 Biomaterials components and their relationship in caArray

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 83. 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 91.

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*).

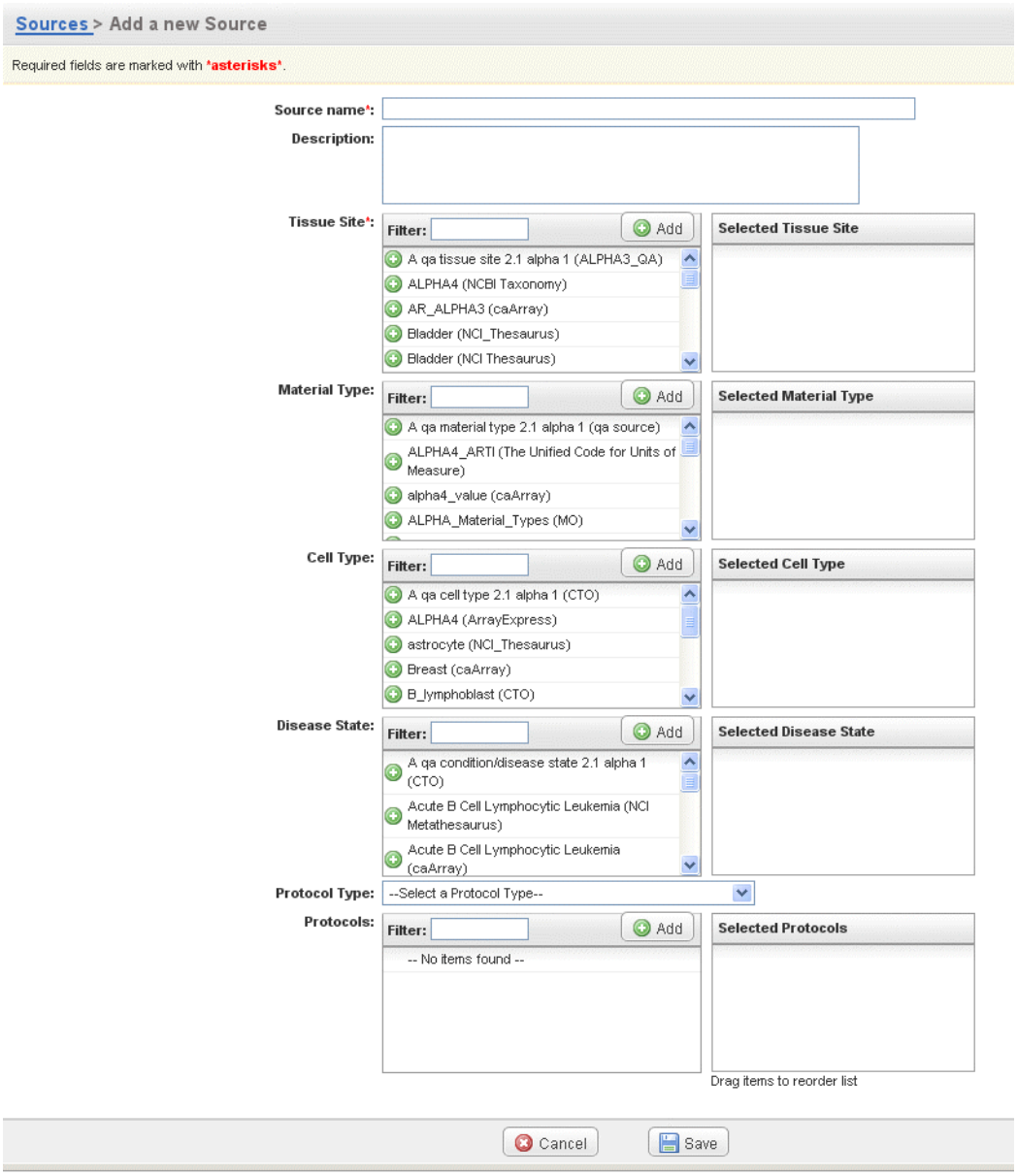


Figure 4.6 Sources subtab

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

Table 4.6 Fields for documenting a source

Source Fields	*Description*
Description	Description of the source
External ID	Enter an additional identifier for the source, beyond the source name. You can create this ID here or you can add it as a field in MAGE-TAB SDRF using a column called Characterisitics [ExternalId]. For more information, see *Importing MAGE-TAB Data* on page 87.
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 *Adding Vocabulary for Experiments* on page 50.
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 *Adding Vocabulary for Experiments* on page 50 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 *Adding Vocabulary for Experiments* on page 50.
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 *Adding Vocabulary for Experiments* on page 50.

Table 4.6 Fields for documenting a source (Continued)

3. 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.

4. Repeat steps 1-3 as often as necessary to add all appropriate sources to the experiment.

5. Proceed to the **Samples** Tab.

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

5. Fields with a red asterisk * are required.

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.

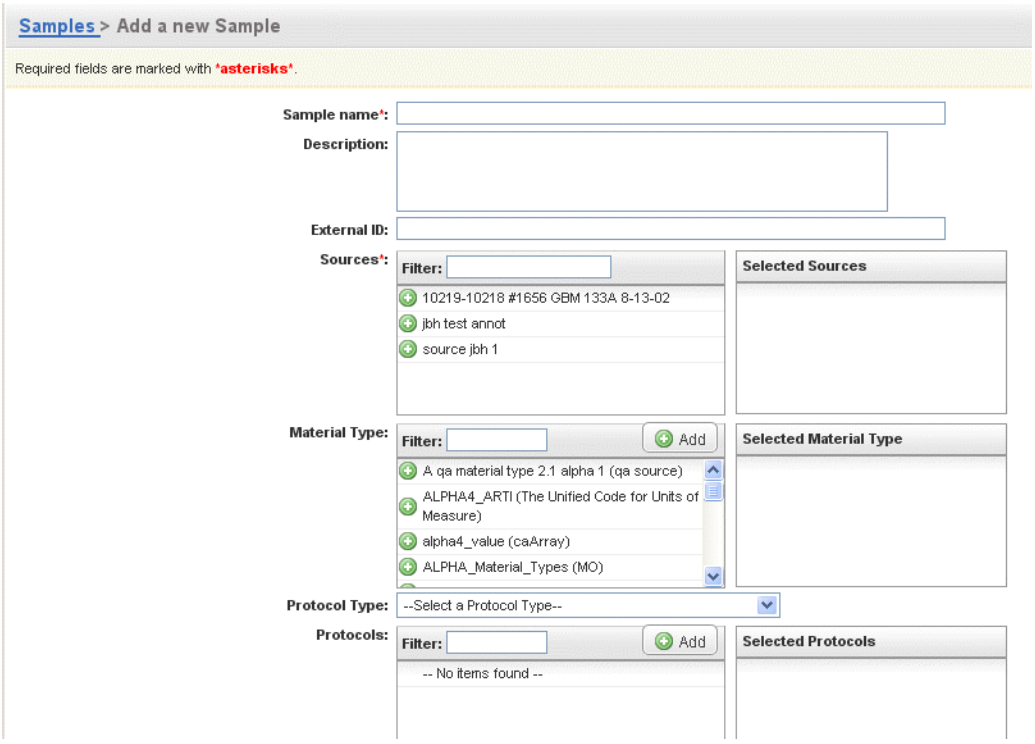


Figure 4.7 A portion of a Samples page

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 for the sample, beyond the sample name. You can create this ID here or you can add it as a field in MAGE-TAB SDRF using a column called Characterisitics [ExternalId]. For more information, see *Importing MAGE-TAB Data* on page 87. 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)*	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. 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 *Sources Tab* on page 37 for more information.
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 *Adding Vocabulary for Experiments* on page 50 for more information about using this feature.
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 (*Managing Protocols* on page 68) or via MAGE-TAB (*Importing MAGE-TAB Data* on page 87).
Protocol	If the appropriate protocol displays in the list, click the adjoining plus icon ( ) to move it into the **Selected Protocols** panel. **Note:** The available selections are limited based on the protocol type selected above. 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. 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 *Creating a Protocol* on page 70.

Table 4.7 Fields for documenting samples (Continued)

3. 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.

4. Repeat steps 1 - 3 as often as needed to enter all the samples used in this experiment.

5. Proceed to the **Extracts** tab.

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]

Extracts Fields	*Description*
Extract Name*	Name assigned to the extract
Description	Description of the extract
External ID	Enter an additional identifier for the extract, beyond the extract name. You can create this ID here or you can add it as a field in MAGE-TAB SDRF using a column called Characterisitics [ExternalId]. For more information, see *Importing MAGE-TAB Data* on page 87.
Samples*	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. 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 *Samples Tab* on page 39 for more information.
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 *Adding Vocabulary for Experiments* on page 50 for more information about using this feature.
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 (*Managing Protocols* on page 68) or via MAGE-TAB (*Importing MAGE-TAB Data* on page 87).

Table 4.8 Fields for documenting an extract

7. Fields with a red asterisk * are required.

Extracts Fields	Description
Protocol	If the appropriate protocol displays in the list, click the adjoining Plus icon ( ) to move it into the **Selected Protocols** panel. **Note:** The available selections are limited based on the protocol type selected above. 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. 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 *Creating a Protocol* on page 70.

Table 4.8 Fields for documenting an extract (Continued)

3. 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.

4. Repeat steps 1 - 3 as often as needed to enter all the extracts used in this experiment.

5. Proceed to the Labeled Extracts tab.

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.

1. 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*).

2. In the Labeled Extract form that opens, enter the information described in *Table 4.9*[8].

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

Table 4.9 Fields for documenting a labeled extract

8. Fields with a red asterisk * are required.

Labeled Extracts Fields	*Description*
External ID	Enter an additional identifier for the labeled extract, beyond the labeled extract name. You can create this ID here or you can add it as a field in MAGE-TAB SDRF using a column called Characterisitics [ExternalId]. For more information, see *Importing MAGE-TAB Data* on page 87.
Extracts*	Extract(s) from which the labeled extract was derived. Extracts must already have been saved to caArray. 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. 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 *Extracts Tab* on page 42 for more information.
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 *Adding Vocabulary for Experiments* on page 50 for more information about using this feature.
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 (*Managing Protocols* on page 68) or via MAGE-TAB (*Importing MAGE-TAB Data* on page 87).
Protocol	If the appropriate protocol displays in the list, click the adjoining Plus icon ( ) to move it into the **Selected Protocols** panel. **Note:** The available selections are limited based on the protocol type selected above. 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. 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 *Creating a Protocol* on page 70.

Table 4.9 Fields for documenting a labeled extract (Continued)

3. 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.

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*).

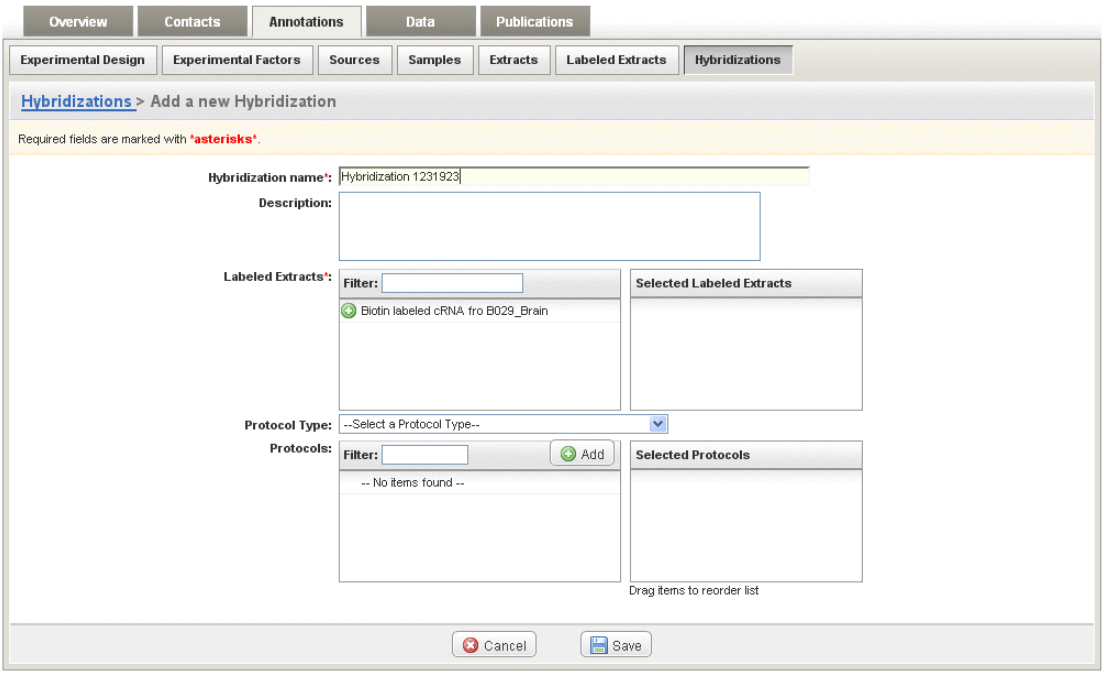


Figure 4.8 Hybridizations page

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

Hybridizations Fields	*Description*
Hybridization Name*	Name assigned to the hybridization

Table 4.10 Fields for documenting a hybridization

9. Fields with a red asterisk * are required.

Hybridizations Fields	*Description*
Description	Description of the hybridization
Labeled Extracts*	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. 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 *Labeled Extracts Tab* on page 43 for more information.
Array Designs	This field displays only if you associated more than one array design on the *Overview Tab*, described on page 31. Select the array design appropriate for this hybridization.
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 (*Managing Protocols* on page 68) or via MAGE-TAB (*Importing MAGE-TAB Data* on page 87).
Protocol	If the appropriate protocol displays in the list, click the adjoining Plus icon ( ) to move it into the **Selected Protocols** panel. **Note:** The available selections are limited based on the protocol type selected above. 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. 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 *Creating a Protocol* on page 70.

Table 4.10 Fields for documenting a hybridization (Continued)

3. 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.

4. Repeat steps 1 - 3 as often as needed to enter all the hybridization data used in this Experiment.

5. Proceed to the *Data Tab* on page 52.

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*).

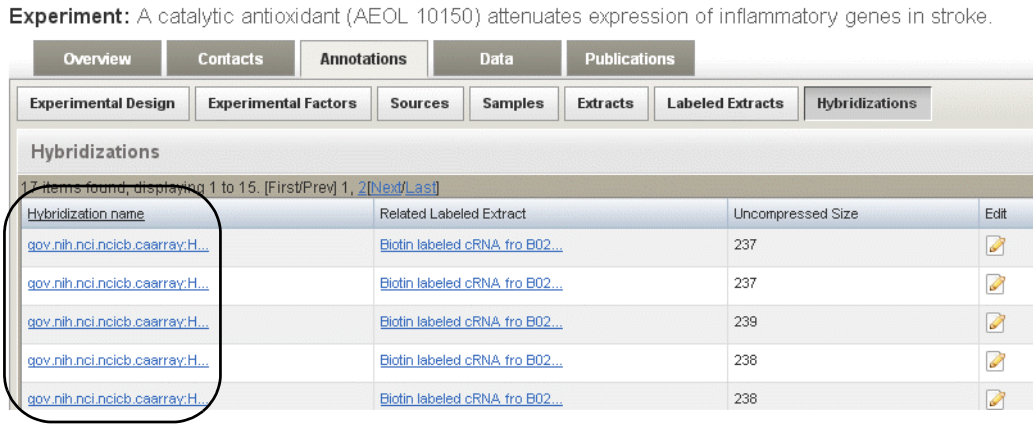


Figure 4.9 Hybridizations associated with an experiment appear on the Hybridizations tab

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

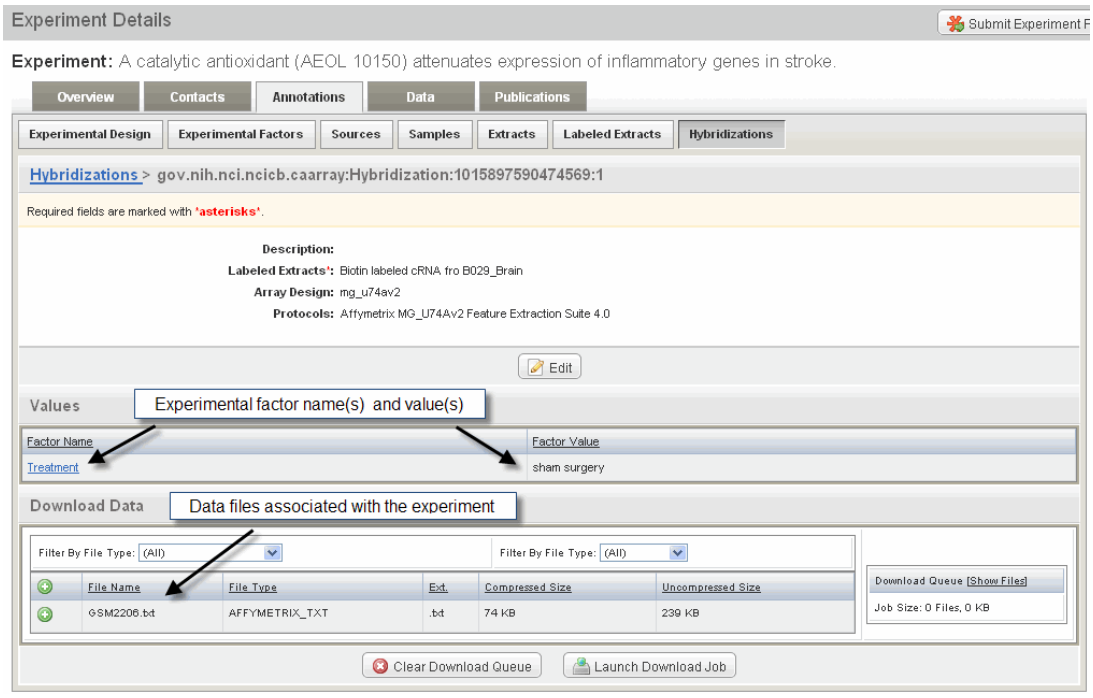


Figure 4.10 Experiment Details, Hybridization page

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 48
- *Copying a Biomaterial/Hybridization* on page 49
- *Deleting a Biomaterial/Hybridization* on page 49
- *Downloading Associated Data Files* on page 49
- *Adding Vocabulary for Experiments* on page 50

Editing Experiment Annotations

As an 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 54.
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*) Note that the copied sample originates from the same source as the original sample..

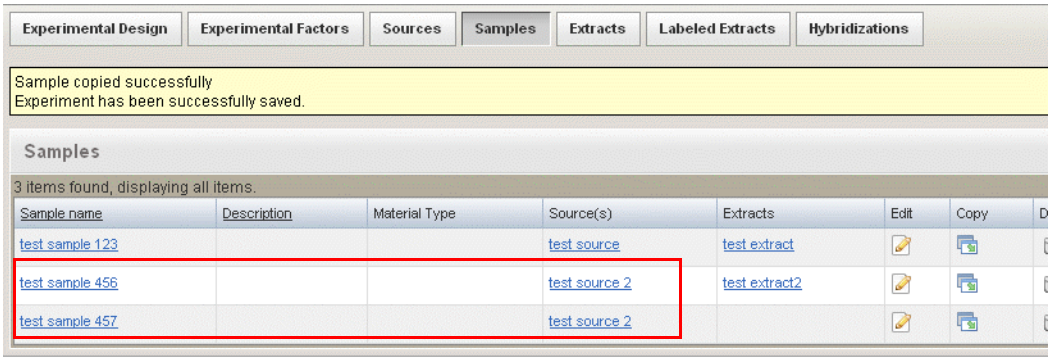


Figure 4.11 A biomaterial or hybridization can be copied in an experiment. The last sample in this list was copied from the previous sample. The name/numeric designation for the sample is automatically generated by caArray.

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 83.

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*).

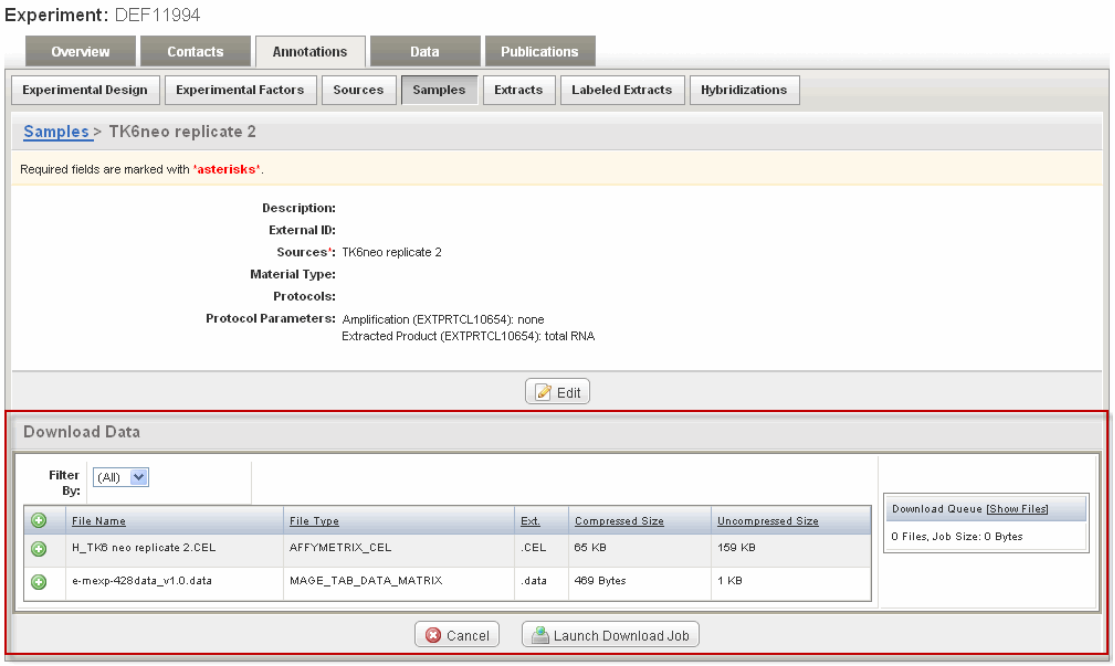


Figure 4.12 Files associated with biomaterials and hybridizations can be downloaded directly from the corresponding tabs.

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 83 and *Downloading Data from caArray* on page 91.
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.

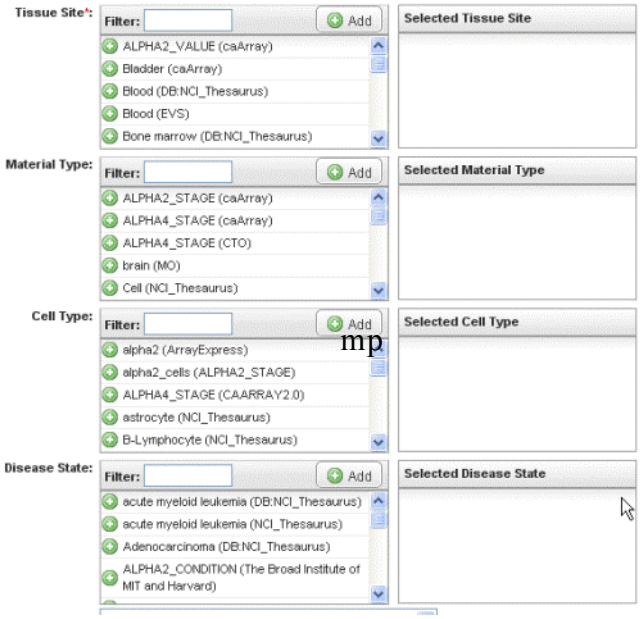


Figure 4.13 Annotation panels for selecting or adding new vocabulary terms to experiments

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 Plus 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. *Example:* DNA
Description	Enter the description of the term, as appropriate. *Example:* 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** • 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. *Example:* http://mged.sourceforge.net/ontologies/MGEDontology.php#DNA
Accession Value	Enter the value given the term in the source vocabulary. *Example:* MO_945

Table 4.11 Fields for entering a new vocabulary term (Continued)

4. 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*.

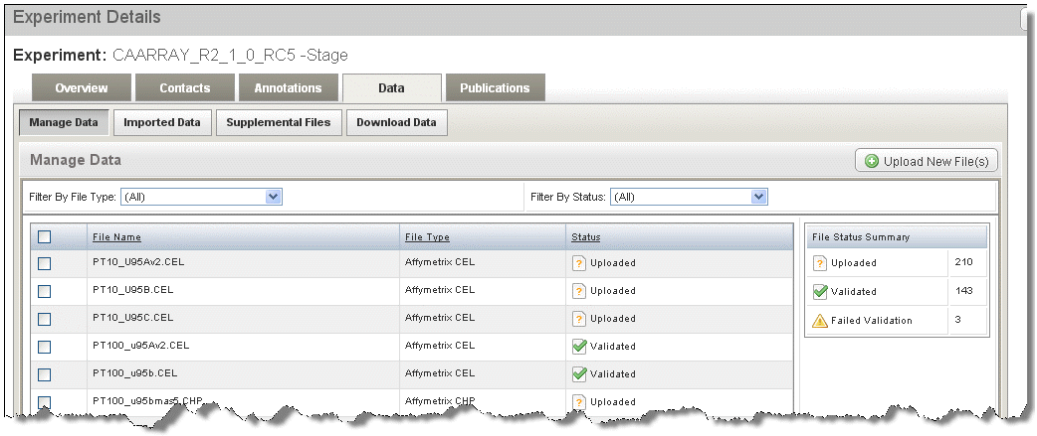


Figure 4.14 caArray experiment data tab, displaying several files that have been uploaded. A status summary displays for all files that have been uploaded into the experiment.

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.

Publications Fields	*Description*
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 (Continued)

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 the "in Progress" state.

You can lock an experiment in the "In Progress" status to make the experiment uneditable by anyone. Click the **Lock Experiment...** button in the upper right corner of the browser window, or the corresponding **Unlock Experiment** button to unlock the experiment.

Note: Locking an experiment does NOT change permissions. You must use the Experiment Permissions page to make the experiment readable by others. See *Experiment Visibility* on page 55.

For information about setting the visibility of an experiment, see *Experiment Visibility* on page 55.

Updating An Experiment Proposal

At any point, after you have saved an experiment draft with In Progress, not Locked, status, you, as its creator and any collaborators can edit it, add or delete data.

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 with In Progress status, you as the experiment creator or any collaborators 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.

   Note: An experiment that has been locked does not have a corresponding Edit button in the Edit column. For more information, see *Experiment Status Settings* on page 54.
3. 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.
4. 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 “In Progress” status (not Locked). 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 your My Experiment Workspace with its status, In Progress. As the experiment creator, once the experiment is In Progress, 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*).

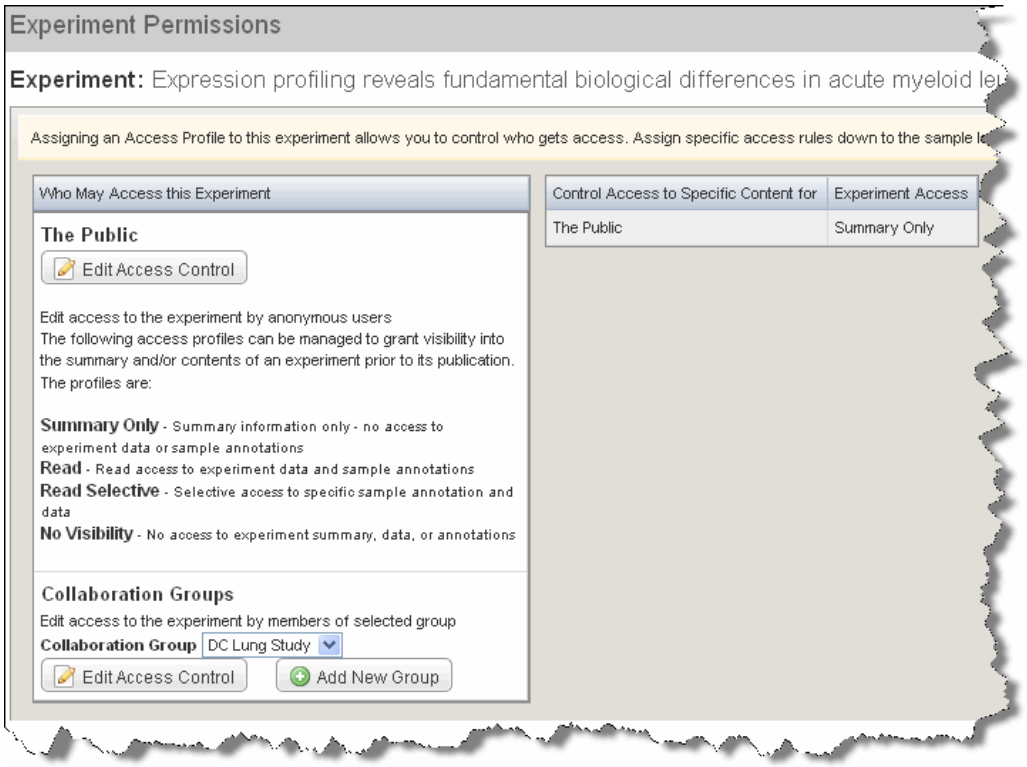


Figure 4.15 Experiment Permissions page

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 on the interface as well as follows:

- **Summary Only** – exposes summary information without access to annotation and array data.
- **Read** – grants read access to the experiment data and sample annotations
- **Read Selective** – grants selective access to specific sample annotations and data. Select this option to apply selective access to experiment samples only. For more information, see *Setting Selective Permissions* on page 56.
- **No Visibility** – allows no access to experiment summary, data or annotations

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*).

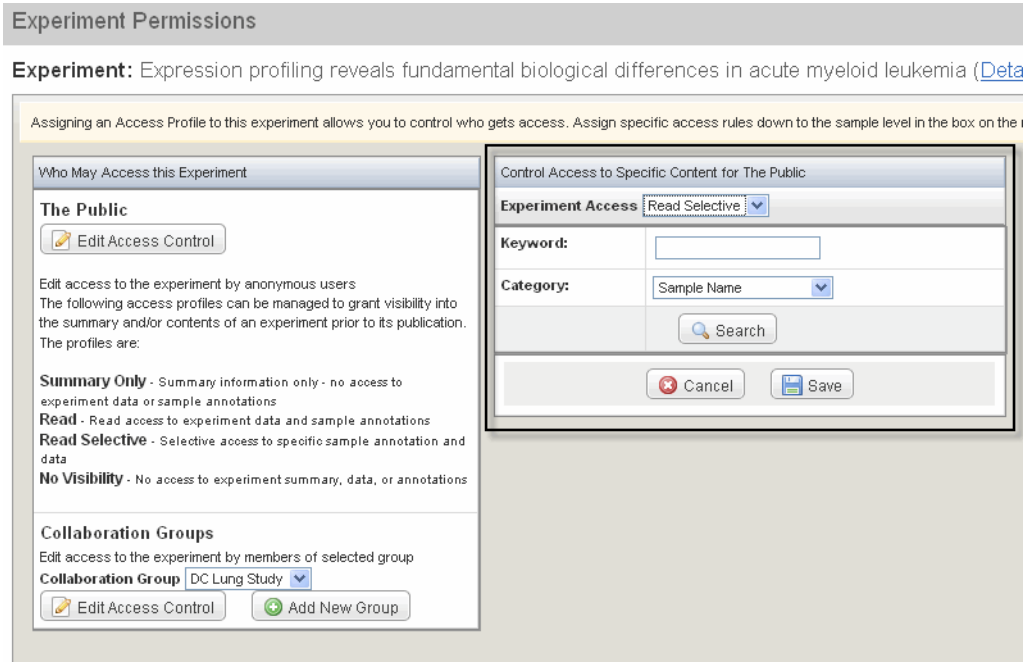


Figure 4.16 Permissions page with Read Selective Access dialog box

3. Click the **Experiment Access** down arrow. This opens the list of visibility options.
4. 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 to be assigned

5. Enter a search keyword in the **Keyword** field.
6. Click the down arrow in the **Category** field, and select the sample metadata criteria for the search.
7. Click the **Search** button to launch the search.

8. 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 visiblity 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 106.

To configure experiment visibility for a collaboration group, 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 for which you want to alter permissions.

   The Experiment Permissions page that opens displays a panel the 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 55.

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 56.
   - **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 106.
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 55.

CHAPTER 5
CURATION TOOLS

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

Topics in this chapter include the following:

Curation Tasks

Curation tasks in caArray 2.4 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*):

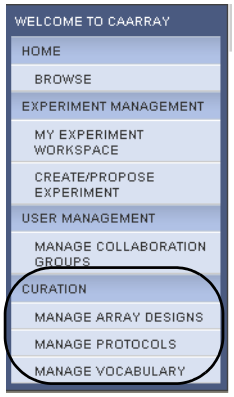


Figure 5.1 Curation options display in the left sidebar

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. You can also download files associated with a selected 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.

File Types	***Validated and parsed as part of the import process***	***Imported but not parsed***
Array Design files	• Affymetrix CDF, PGF CLF • Illumina Design CSV, BGX/TXT • GenePix GAL • Agilent GEML/XML • Nimblegen NDF	• Agilent CSV, XML • UCSF Spot SPT • ImaGene TPL • MAGE-TAB ADF **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. **Note:** caArray may have new parsers available for array design files in the system that are already imported but not parsed. To learn about retrofitting those files, see *Retrofitting Array Design Files* on page 66.

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 83. Properties corresponding to those array designs are described in Table *5.2* .

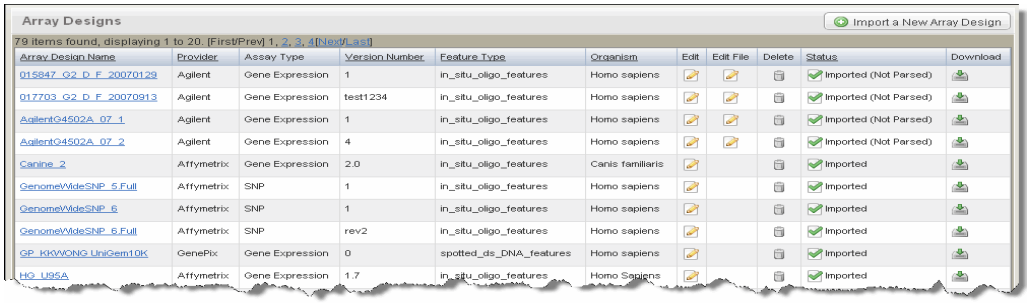


Figure 5.2 Array Designs imported into caArray

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

Array Designs Properties	***Description***
Array Design Name	Name assigned to the array design
Assay Type	The assay type used for the Array Design. • **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
Provider	Select from the drop-down menu the provider of the array. **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 *Managing Data* on page 75.
Version Number	The version number of the array design
Feature Type	The technology type or platform of the reporters on the array. Note that these terms are from the MGED Ontology.
Organism	The organism the array was designed to assay.

Table 5.2 Array Designs properties

Array Designs Properties	*Description*
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 (Continued)

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*).

2. 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.

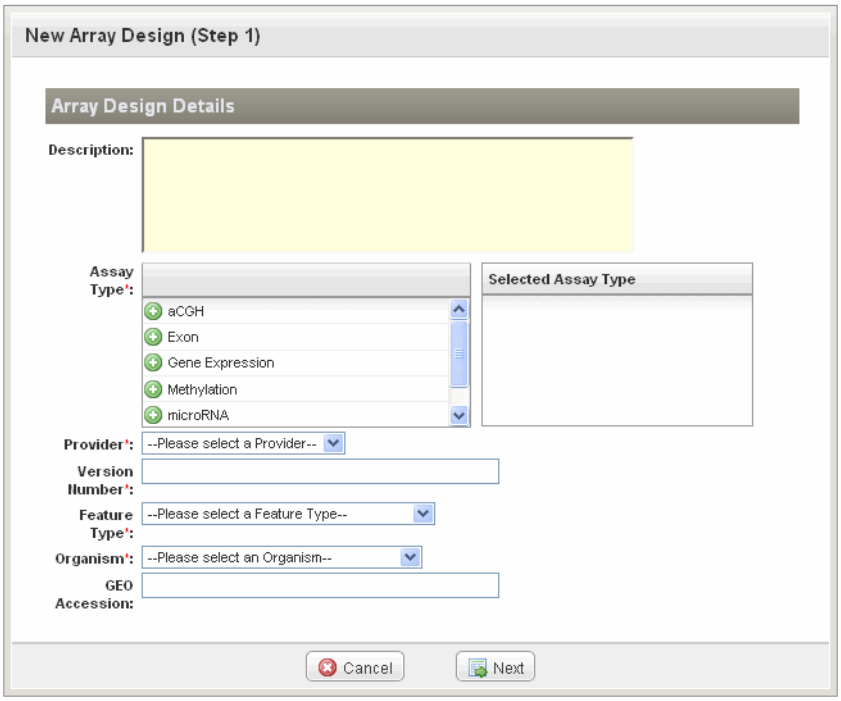


Figure 5.3 Manage Array Designs dialog box

3. 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.
Assay Type*	Select one or more assay types used for the array design. • **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. • **Gene Expression** – experiment using microarrays intended to measure levels of transcribed genes • **Methylation** – experiment that attempts to establish patterns of methylation genome-wide or within targeted promoters or CpG islands • **microRNA** – Experiment that measures activity among the 217 genes encoding miRNA. Patterns of gene activity that can distinguish types of cancers can be discerned. • **SNP** – experiment using microarrays intended to detect nucleotide changes in chromosomal DNA
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.
GEO Accession	**Note:** This value is pre-populated for common array designs that are in the GEO repository. This is done at the NCI CBIIT instances of caArray and at local installations as part of the upgrade process to v.2.4. For an array design for which the GEO accession is not populated, enter the information here.

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.

11. Fields with a red asterisk * are required.

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, 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.

Retrofitting Array Design Files

Any array design files that were imported but not parsed in caArray prior to v.2.4 can be eligible for parsing with appropriate new parsers in v.2.4.

To learn if your data files are eligible for retrofitting, see the following:

- If you are logged in as a system admin, follow these steps.

a. Note the “TO DO” category on the left sidebar that displays a RE-PROCESS menu option (*Figure 5.4*).

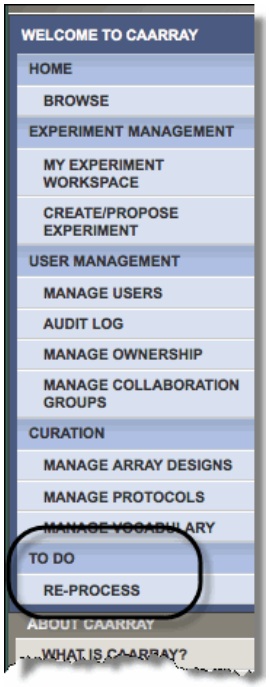


Figure 5.4 Sidebar option for beginning retrofitting process

b. Click **RE-PROCESS** to open a Manage New Imports page. Select the Array Designs tab that lists all array designs that are eligible for retrofitting or parsing.

c. Select an array design you want to reprocess and click the **Reparse** button ( ).

- If you are logged in as a user, open the Manage Array Designs page.

  a. Select an eligible array design and re-import it. See *Adding an Array Design* on page 64 for more information.

The “imported, not parsed” array design will be parsed and the appropriate probes and/or features will be stored in the repository. If the import succeeds, the array design will be in the **Imported** state. If the import fails, it will be in the **Import Failed** state.

Deleting an Array Design

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

- 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.

Downloading Files Associated with Array Design

On the Manage Array Designs page, click the **Download** icon ( ) corresponding to an array design of your choice. This opens an Opening... dialog box that displays the file name. Select the option to either open or to save the file to your local drive.

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 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 5.5*).

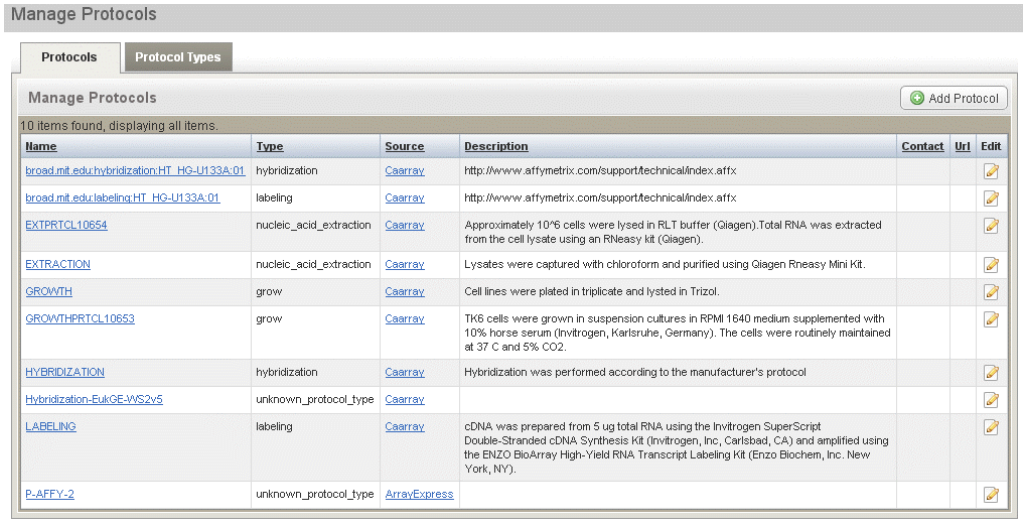


Figure 5.5 Protocols page

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***
Name	Name assigned the protocol

Table 5.4 Protocol properties

Protocol Properties	*Description*
Type	Descriptor of the protocol type, such as labeling or hybridization.
Description	Description of the protocol procedure. Include any and all appropriate details, such as the detailed steps taken in a laboratory procedure. *Example*: Enter a description of a procedure for labeling RNA with fluorescent tags to be used in a hybridization procedure.
Contact	The name of the person to contact for information about the protocol.
URL	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 *Editing a Protocol* on page 71.

Table 5.4 Protocol properties

Tip: 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.

Table 5.5 Protocol properties

Protocol Type Properties	Description
Edit	Click the **Edit** icon ( ) to open the protocol type details page where you can edit the data. For more information, see *Editing a Protocol Type* on page 71.

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 5.5*).

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
Description	Description of the protocol procedure. Include any and all appropriate details, such as the detailed steps taken in a laboratory procedure. *Example*: Description of a procedure for labeling RNA with fluorescent tabs to be used in a hybridization procedure.
Type*	Descriptor of the protocol type such as "labeling" or "hybridization" from a controlled vocabulary, for example MGED. If the appropriate protocol displays in the list, click the adjoining Plus icon ( ) to move it into the **Selected Protocols** panel. **Note:** The available selections are limited based on the protocol type selected above. 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. If the appropriate protocol has not been entered into the system, click **Add** to open the page where you can add a new protocol.

Table 5.6 Protocol fields

12. Items with an asterisk are required.

Protocol Properties	*Description*
Contact	The name of the person to contact for information about the protocol.
Software	Name of software used in the protocol. *Example*: GenePix Pro 3.0.1.22
Hardware	Name of hardware used in the protocol. *Example*: GeneChip® Fluidics Station 450®
URL	Link to a source of external documentation related to the protocol

Table 5.6 Protocol fields

4. 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:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Protocols.**
2. 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 (Edit) 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 (Edit) 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 5.6*).

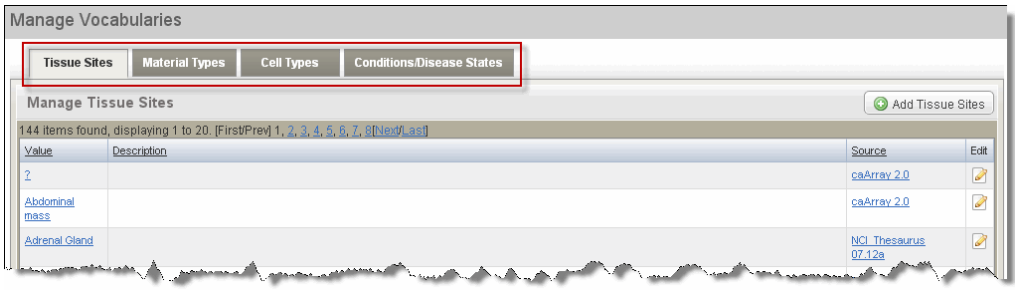


Figure 5.6 Manage Vocabularies page

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***
Value	Name assigned the term
Description	Description of the vocabulary term
Source	The external source or standardized vocabulary that is the source of the term.

Table 5.7 Protocol properties

Protocol Properties	Description
Edit	Click the **Edit** icon ( ) to open the vocabulary details page where you can edit the information for the term. For more information, see *Editing a Vocabulary Term* 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 50) or you can work with vocabulary terms using on 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. *Example:* DNA
Description	Enter the description of the term, as appropriate. *Example:* deoxyribonucleic acid
SOURCE	
Create a New Source [for the Term you are adding]	Select **Yes** or **No** • 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	

Table 5.8 Fields for entering a new vocabulary term

Vocabulary Term Category	*Description of Fields*
Accession URL	Enter the exact URL for accessing the new term. *Example:* http://mged.sourceforge.net/ontologies/MGEDontology.php#DNA
Accession Value	Enter the value given the term in the source vocabulary. *Example:* MO_945

Table 5.8 Fields for entering a new vocabulary term

4. 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:

1. After logging into caArray, on the left sidebar under **Curation**, click **Manage Vocabulary.**
2. Select the attribute tab where the term you want to edit is listed.
3. 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 (Edit) 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 chapter describes the processes for submitting data such as annotation and array content into caArray experiments.

The following topics are part of this chapter:

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 62.

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 more comprehensive and specific information on this topic, see *Appendix B, Importing Data Files*.

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 89.

File Types	***Imported after validation and parsing***	***Imported without validation and parsing***
Raw/processed data files	• Affymetrix CEL, CHP, CNCHP • GenePix GPR* • Illumina CSV, Sample Probe Profile TXT, Genotyping processed data matrix TXT, Raw TXT • Agilent Raw TXT • Nimblegen Normalized Pair Report TXT *For more information about GenePix GPR and Illumina CSV files, see *MAGE-TAB SDRF Validation Rules* on page 120, items *1.* and	• Affymetrix DAT, RPT, TXT, and EXP • Agilent TSV, derived TXT • Illumina IDAT, TXT • ImaGene TIF, TXT • Nimblegen GFF, Raw or Derived TXT • ScanArray CSV • GEO SOFT, GSM **Note:** caArray may have new parsers available for data files in the system that are already imported but not parsed. To learn about retrofitting those files, see *Retrofitting Data Files* on page 86.

Table 6.1 File types that cam be imported into caArray

File Types	Imported after validation and parsing	Imported without validation and parsing
Array Design files	• Affymetrix CDF, PGF, CLF • Illumina Design CSV • Genepix GAL **Note:** These can be uploaded, validated and imported only through the Manage Array Design feature described in *Managing Array Designs* on page 62.	• Agilent CSV, XML • UCSF Spot SPT • ImaGene TPL • Nimblegen NDF • GEO GSM **Note:** These can be uploaded, validated and imported only through the Manage Array Design feature described in *Managing Array Designs* on page 62. **Note:** caArray may have new parsers available for array design files in the system that are already imported but not parsed. To learn about retrofitting those files, see *Retrofitting Array Design Files* on page 66.
MAGE-TAB files	• MAGE-TAB SDRF (Sample and Data Relationship Format) • MAGE-TAB IDF (Investigation Description Format) only, no referenced SDRFs • MAGE-TAB Copy Number Data Matrix **Note:** Only one IDF is allowed per import, since the import is in the context of a single experiment.	• MAGE-TAB ADF • MAGE-TAB Data Matrix (not copy number)

Table 6.1 File types that cam 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 113 as well as *Appendix B, Importing Data 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 *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.
Supplemental Files	This tab lists files and documents that have been uploaded to caArray, and identified as supplementary (reference) files.
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.

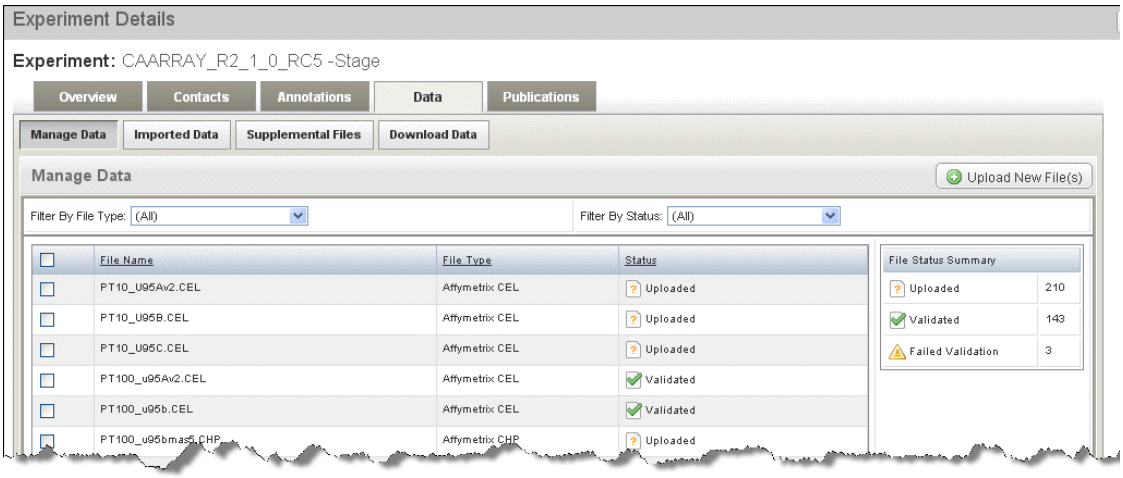


Figure 6.1 Data tab displays much information about uploaded files

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 62.

Note: It is not possible to import source or sample annotations directly into caArray 2.4 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 83. You can download data files from the biomaterials tabs, which indicate associations with sources, samples, extracts, labeled extracts and/or hybridization.

Uploading Data Files

Through the process of uploading annotation and array data, the content becomes available for validation and import into caArray. You can share imported files for download or you can delete them.

Notes:

- Individual files or multiple files packaged in a .zip format can be uploaded. Due to browser limitations, the combined size of your upload must be less than 2 GB. If you need to upload more data, you must do so in multiple steps.
- An individual file (whether standalone or packaged in a zip) can only be half as large as the amount of memory on your server for the validation and import processing to occur. caArray 2.4 has been tested using a 2 GB allocation of memory and therefore the maximum size of any individual, un-zipped file is 1 GB.
- caArray supports the upload of .zip compressed files only. NO other compression formats are supported in v.2.4 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

Note: You can upload data only for experiments for which you have appropriate permissions. See *Experiment Visibility* on page 55.

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*). .

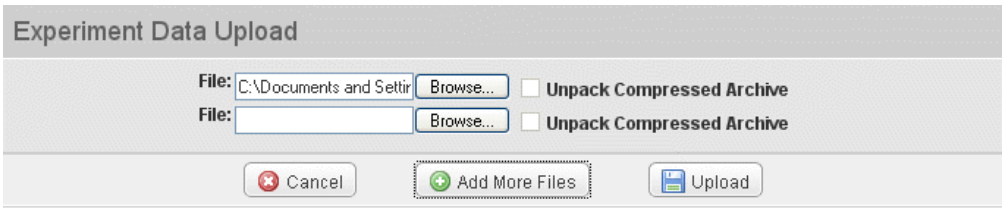


Figure 6.2 Experiment Data Upload dialog box

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 informs 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*).

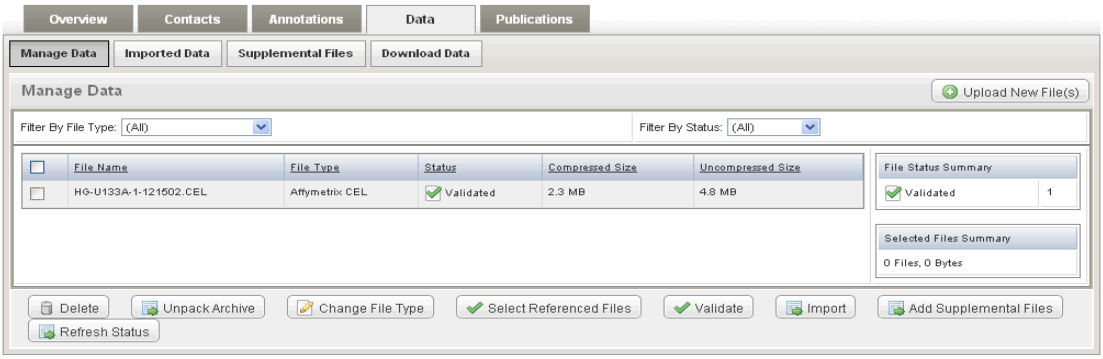


Figure 6.3 The Manage Data subtab displays files in caArray and their status. The first listed file in "In Queue" to be validated

Once files are uploaded, the files workflow should be continued 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.

Selecting Referenced Files

If you have uploaded an IDF file and it is visible on the Manage Data tab, you can easily select a set of related files and validate or import the set as a whole. Select the button next to the IDF file, then click **Select Referenced Files**. caArray selects the SDRF file referenced by the IDF as well as any data files referenced by the SDRF. (The files must have already been uploaded into caArray.)

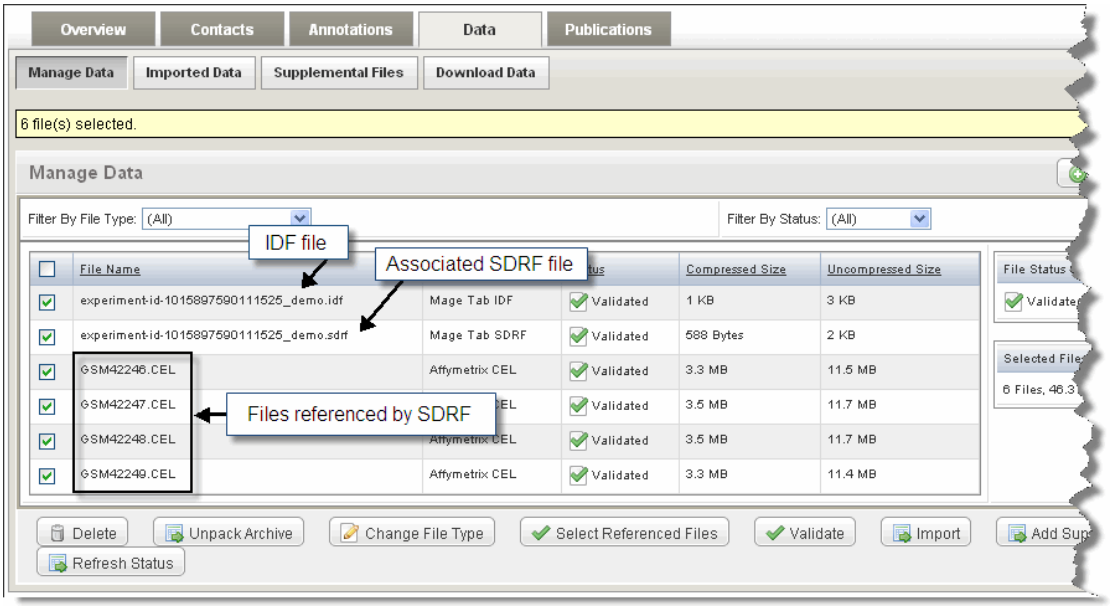


Figure 6.4 A set of files referenced by an IDF and its associated SDRF file can be selected and validated and imported as a whole.

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: You can delete imported files 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. You can also delete data files with an "Imported, not Parsed" status. This deletes the files themselves, but not the biomaterials and hybridizations with which they are associated.

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.

Validating Data Files

Once data have been uploaded into caArray, anyone with writing 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. For more comprehensive and specific information on this topic, see *Appendix B, Importing Data Files*.

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 routinges are available. You should identify these as supplemental files. Instructions are provided in *Supplemental Files* on page 89.

- 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 76.

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 (

). 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**.

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

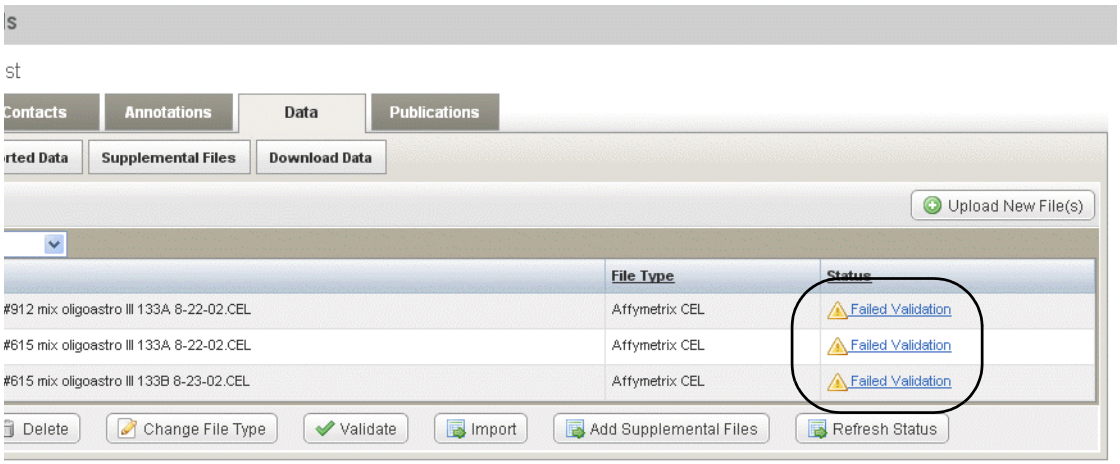


Figure 6.5 Validation failures display on the Manage Data table.

Importing Data

Note: This section 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 87 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 is created, and all of the files are 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 87 as well as *Appendix B, Importing Data Files*.

Files can be downloaded from the Data tab or from the Annotations tabs. For more information, see *Downloading Data from caArray* on page 91.

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.

caArray may have new parsers available for data files in the system that are already imported but not parsed. caArray provides the ability to "retrofit" those files. For more information, see *Retrofitting Data Files* on page 86.

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 81.
  - 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:
   a. **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`.

b. **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.

c. **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**.

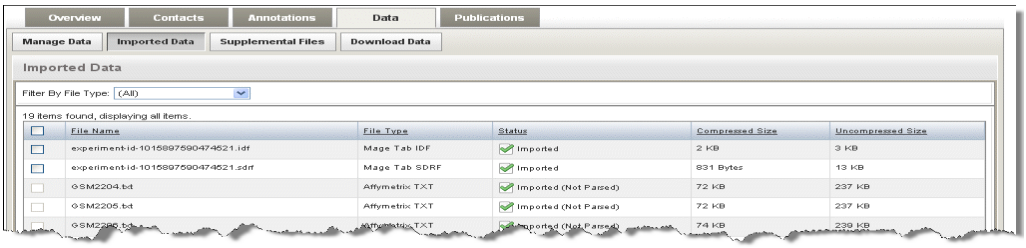


Figure 6.6 Imported Data tab displaying imported data files

7. To filter the data display by file type, click the **Filter by File Type** drop-down arrow.

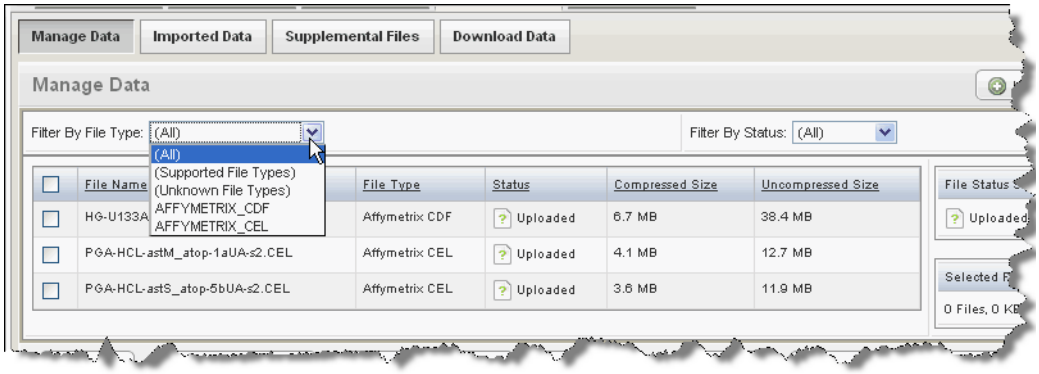


Figure 6.7 Drop-down list for filtering imported files by file type

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

Retrofitting Data Files

Any data file that was imported but not parsed in caArray prior to v.2.4 can be eligible for parsing with appropriate new parsers in v.2.4.

To learn if an existing data file is eligible for parsing ("retrofitting"), see the following:

- If you are logged in as a system admin, follow these steps:

  a. Note the "TO DO" category on the left sidebar that displays a RE-PROCESS menu option (*Figure 6.8*).

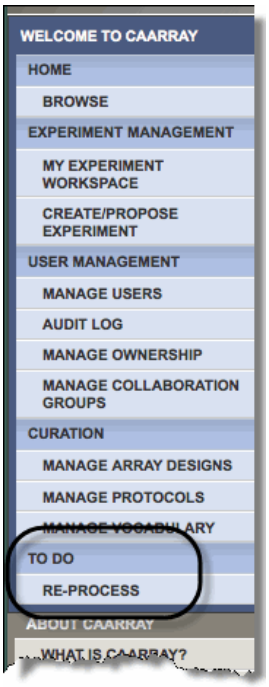


Figure 6.8 Sidebar option for beginning retrofitting process

  b. Click **RE-PROCESS** to open a Manage New Imports page.

c. The Experiments tab lists all experiments that have existing data files that are eligible for retrofitting or parsing. Click on the **Experiment ID** link for the files you want to reprocess.

d. On the Experiment Details page, Data/Imported Data tabs, check the file you want to reprocess and click the **Reparse** button (*Figure 6.9*).

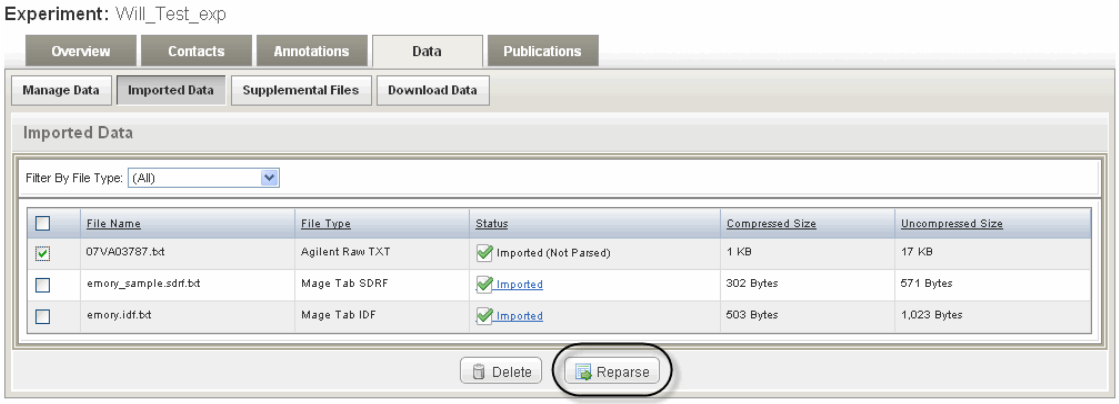


Figure 6.9 Data tab/Imported Data tab showing checked unparsed file and the reparse button

- If you are logged in as a user, follow these steps:

  a. Navigate to your Experiment Workspace which shows the experiments for which you are the data owner. In the Reparse column, caArray indicates experiments that have array design or data files eligible for reparsing.

  b. Click the Experiment ID link which opens the Experiment Details page.

  c. Open the **Data** tab > **Imported Data** page for the experiment.

  d. Select an eligible data file and click the **Reparse** button.

After selecting any of these reparsing options, caArray parses the data file, validates it against the array design and creates the parsed data values. If the reimport succeeds, the status is shown as Imported. It the file cannot be parsed due to errors, the file remain in the IMPORTED NOT PARSED state. View the error log for details.

See *Retrofitting Array Design Files* on page 66 for information. about retrofitting array design files.

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 83.

For caArray-specific information regarding MAGE-TAB, see *Appendix A*, *MAGE-TAB in caArray*.

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 (Minimum Information about a Microarray Experiment) 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 2.4, 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). 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 79
- *Validating Data Files* on page 81
- *Importing Data* on page 83
- *Appendix A, MAGE-TAB in caArray*, on page 111

MAGE-TAB Import with Existing Experimental Components

If the MAGE-TAB data set you are importing refers to biomaterials, hybridizations, data files or experimentatl 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 during import. You can add new characterisitcs during import, as well. 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 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 origianlly 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.

For more information, see *Auto-Generated Missing Biomaterials in MAGE-TAB* on page 112 and *Protocols Associated Intelligently* on page 112.

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 91.

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 **Add Supplemental File** button.

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

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

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 87. 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 83.

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.

The Download Data page has three sub-menus, shown in *Figure 7.1*..

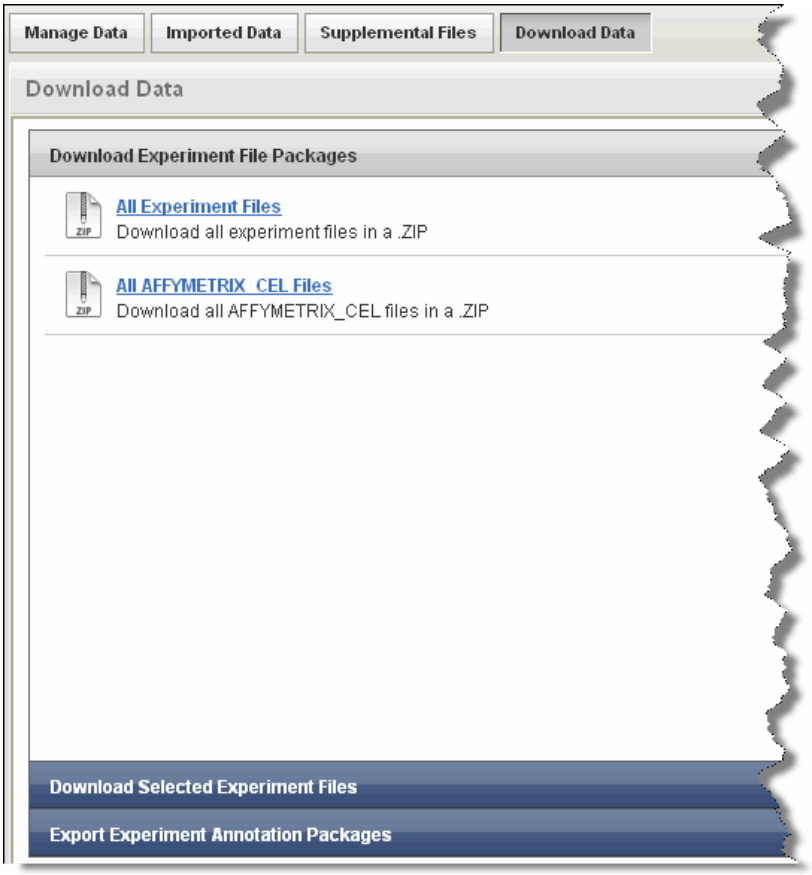


Figure 7.1 Download Data tab

For information about each, follow the appropriate link:

Downloading Selected Experiment Files

Click **Download Selected Experiment Files** to select specific files for download. 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 7.2*).

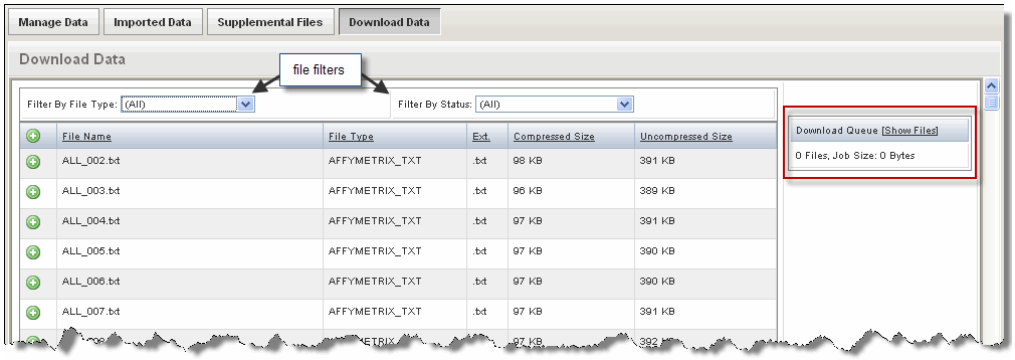


Figure 7.2 Download data subtab

1. To filter the list of files on the tab, click one of the following drop down menus:
   a. Click the **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.
2. 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 2.0 GB (after compression), the archive is downloaded in .tar.gz format.
3. To remove selected files from the queue, click the **Remove** icon(s) ( ) corresponding to the data file or click the **Clear Download Queue** button.
4. 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.
5. 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.

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

Exporting Experiment Annotation Packages

The **Export Experiment Annotation Packages** option allows you to export the experiment annotations (and optionally data) into MAGE-TAB and GEO SOFT formats.

Exporting to MAGE-TAB

If you export to MAGE-TAB, 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 **MAGE-TAB (Annotations Only)** link.

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.

Exporting to GEO SOFT

Note: The export to GEO SOFT option is available only for Affymetrix experiments in caArray v.2.4.

- If you choose to export to GEO SOFT, (annotation file only), caArray validates that all manadatory annotations are present, translates the experiment annotations into GEO SOFT format and prompts you to save the generated .soft.txt file.
- If you export to the GEO SOFT submission package (annotation file and data files), caArray validates that all mandatory annotations are present, translates the experiment annotations into GEO SOFT format and prompts you to save an archive containing the .soft.txt file, all raw, derived and supplementary files, and a README file that explains how to submit to the GEO repository, if you so choose. If the archive is less than 2 GB, the package is provided in the .zip format; otherwise, it is provided in tar.gz format.

Validation Details for GEO SOFT Export:

- The experiment must have at least one array design.
- The provider for the array designs must be Affymetrix.
- All array designs associated with the experiment must be ones for which caArray has the GEO accession.
- Every hybridization must have at least one raw data file.
- Every hybridization must have a derived data file of type AFFYMETRIX_CHP.
- The following protocols must be present: extract, label, hybridization, scan and data processing. (These are entered via MAGE-TAB.)
- There must be at least one characteristic or factor value that is present for every hybridization somewhere in its biomaterial chain.
- For every biomaterial chain, the Material Type of the extract or labeled extract must be present.
- For every biomaterial chain, the label of the labeled extract must have a non-empty, non-null value.
- caArray informs you of the result of the validation, and provides the list of errors if the validation did not succeed.

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 2.4.

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 in caArray. It also discusses the processes for managing ownership and collaboration group access to experiments 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:

Administering caArray User Accounts Using UPT

Note: If you are interested in registering for a caArray account, see *Registering as a New caArray User* on page 10.

In caArray 2.3, all tasks related to creating and managing user accounts can be performed only by a System Administrator using the CBIIT User Provisioning Tool (UPT) 4.2.The following section discusses the use of the UPT for performing these tasks. For further information about UPT, see Chapter 3 of the user's guide located here: https://gforge.nci.nih.gov/docman/view.php/12/18945/caCORE_CSM_v42_ProgrammersGuide.pdf

Relationship Between caArray and UPT

The User Provisioning Tool (UPT) is a separately installed application which serves as the user management interface for all National Cancer Institute CBIIT Life Sciences Distribution (LSD) applications, caArray included. The UPT application is the central point for all user management functionality within caArray. Whenever a new user needs to be granted access to caArray and be assigned to predefined groups (like

"Principal Investigator, "Lab Scientist" etc.), UPT is the application that is used to do this. Because UPT is the user management tool for all LSD applications, it is not caArray-specific and you need to apply some caArray-application related configurations prior to using it for caArray user administration. After proper configuration, you can use UPT to add new users, and apply user group assignments to the caArray database directly. Note that "groups" in UPT are different from "collaboration groups" in caArray. The UPT "groups" refer to predefined groups like Principal Investigator or Lab Scientist, which determine what roles the user has. (See *Roles in caArray* on page 105.) The caArray "collaboration groups" are a way to group multiple users (for example, into a "UPenn collaboration group") and control their access to experiments. UPT "groups" are managed within UPT, and caArray "collaboration groups" are created and managed within the caArray application. (See *Managing Experiment "Ownership" and Group Access* on page 106.)

UPT Users

When using UPT, there are two users of interest:

- **superadmin:** This is the user that administers all applications inside UPT. They are responsible for creating and configuring applications inside UPT, as well as creating and assigning application administration users, such as the caArray system administrator discussed immediately below. If using database authentication configuration for UPT and caArray, this user's login ID is "superadmin", but if using LDAP authentication configuration for UPT and caArray, this user's login ID should match the `uid` value for the LDAP user to be assigned UPT administration responsibilities. To do this, the values of these properties must be changed in the UPT `project.properties` file for the command line installer prior to UPT installation:

  `super.admin.user`: uid for LDAP user to administer UPT.

  `super.admin.first.name`: The first name for user above.

  `super.admin.last.name`: The last name for the user above.

  Unfortunately, the UPT GUI installer does not currently support configuration of these properties, so a manual DB update must be done after installing UPT. For example, to set the correct UPT superadmin user parameters for a hypothetical user named John Doe, you can execute the following SQL:

  ```
  UPDATE CSM_USER SET LOGIN_NAME = 'jdoe', FIRST_NAME = 'John',
  LAST_NAME = 'Doe'  WHERE LOGON_NAME  = 'superadmin';
  ```

- **caArray system administrator:** This is the user who will be logging in to UPT to administer the caArray application, creating new caArray users and assigning them to appropriate groups.

caArray Application Configuration in UPT

To use UPT to administer caArray users, first the UPT superadmin user must create a caArray application configuration inside UPT.

To create a caArray application configuration, follow these steps:

1. On the UPT application homepage, log in as as superadmin user, providing the following superadmin user information (default credentials, which can be changed from the USER tab within UPT itself):
   - **LOGIN ID:** superadmin (or appropriate LDAP uid)
   - **PASSWORD:** changeme (or appropriate LDAP password)
   - **APPLICATION NAME:** csmupt

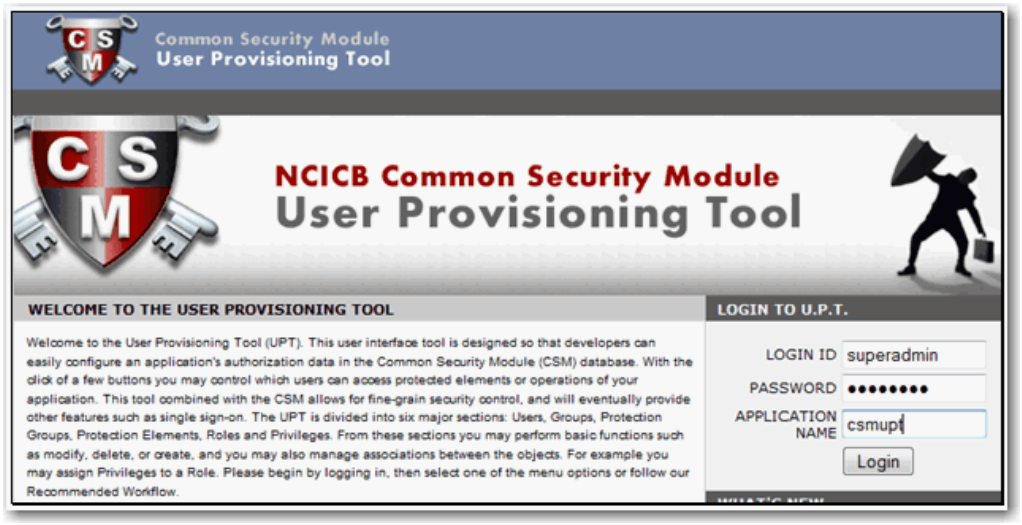


Figure 8.1 UPT home page which is also the login page

2. After logging in as the UPT superadmin user, select the APPLICATION tab, and then create a new application, caArray, named **exactly** (case-sensitive) as that specified as the CSM application name during caArray installation (default is “caarray”). For more information, see the caArray documentation on the NCI wiki: https://wiki.nci.nih.gov/x/YBohAQ.
3. Specify the database connection information so that UPT can make database updates directly.

   Tip: If you have already installed caArray, you can obtain all of the database connection information from the `<caArray installation home>/jboss-4.0.5.GA/server/default/deploy/caarray-mysql-ds.xml` file.

   You must specify the following information:
   - **Application Database URL:** `<connection-url>` element contents from the `caarray-mysql-ds.xml` file.
   - **Application Database User Name:** `<user-name>` element contents from the `caarray-mysql-ds.xml` file.
   - **Application Database Password:** `<password>` element contents from the `caarray-mysql-ds.xml` file.
   - **Application Database Dialect:** `org.hibernate.dialect.MySQLDialect`

- **Application Database Driver:** `com.mysql.jdbc.Driver`
- **Application CSM Version:** You must enter ***4.1***.

4. Click **Add** to submit the details.

   Figure 8.2 shows an example of a successfully executed application.

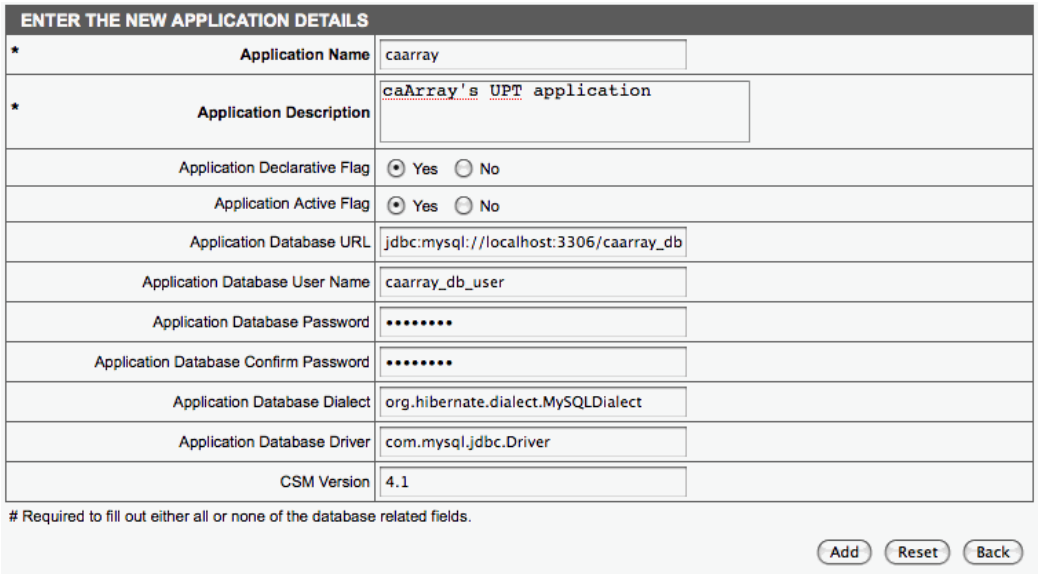


Figure 8.2 UPT form for entering caArray application information

Creating an Admin User

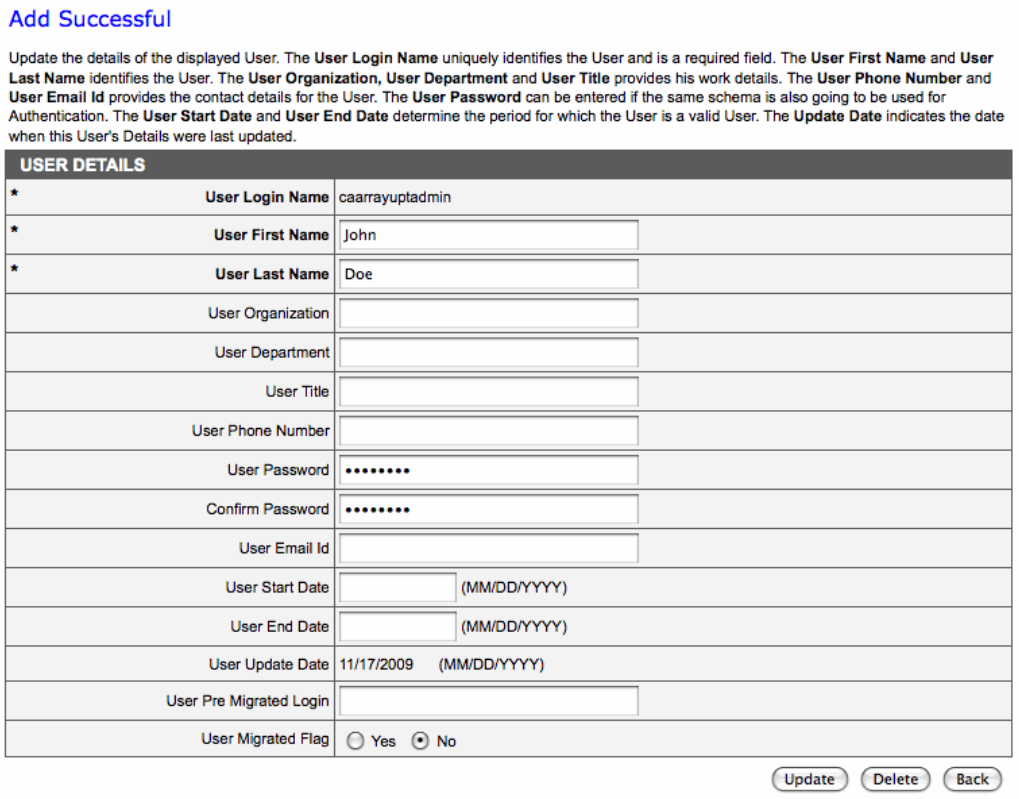
1. After successfully configuring the application, you must create a user who will administer the caArray application within UPT. Select the USER menu option, and select the **Create a New User** link.

2. Enter details for the following required fields:

   - **User Login Name**
   - **User First Name**
   - **User Last Name**
   - **User Password** (If you are naming an LDAP user, do not specify the password.)

3. Click **Add** to confirm the new user.

Figure 8.3 shows an example of a successfully created user.



Add Successful

Update the details of the displayed User. The **User Login Name** uniquely identifies the User and is a required field. The **User First Name** and **User Last Name** identifies the User. The **User Organization, User Department** and **User Title** provides his work details. The **User Phone Number** and **User Email Id** provides the contact details for the User. The **User Password** can be entered if the same schema is also going to be used for Authentication. The **User Start Date** and **User End Date** determine the period for which the User is a valid User. The **Update Date** indicates the date when this User's Details were last updated.

USER DETAILS	
* **User Login Name**	caarrayuptadmin
* **User First Name**	John
* **User Last Name**	Doe
User Organization	
User Department	
User Title	
User Phone Number	
User Password	••••••••
Confirm Password	••••••••
User Email Id	
User Start Date	(MM/DD/YYYY)
User End Date	(MM/DD/YYYY)
User Update Date	11/17/2009 (MM/DD/YYYY)
User Pre Migrated Login	
User Migrated Flag	○ Yes ◉ No

Update Delete Back

Figure 8.3 UPT form for entering caArray UPT user information

4. After successfully creating the account, you must associate this user with the caArray application as an administrator. On the APPLICATION tab click **Select an Existing Application.**

5. Enter the caArray application name in the **Application Name** text field. (Use “caarray”, unless you changed it during caArray installation.) Click **Search.**

   UPT should display the caArray application in the SEARCH RESULTS.

6. Select the **caArray** application and click **View Details**.

7. Click the **Associated Admins** button at the bottom of the APPLICATION DETAILS form, then click the **Assign Admin** button on the Application And Admin Association page.

8. In the pop-up dialog, specify the **User Login** Name of the user account that you created previously, and then click the **Search** button.

9. Select the user and click the **Assign Admin** button. In the page that opens, the user should appear in the Assigned Administration section (*Figure 8.4*).

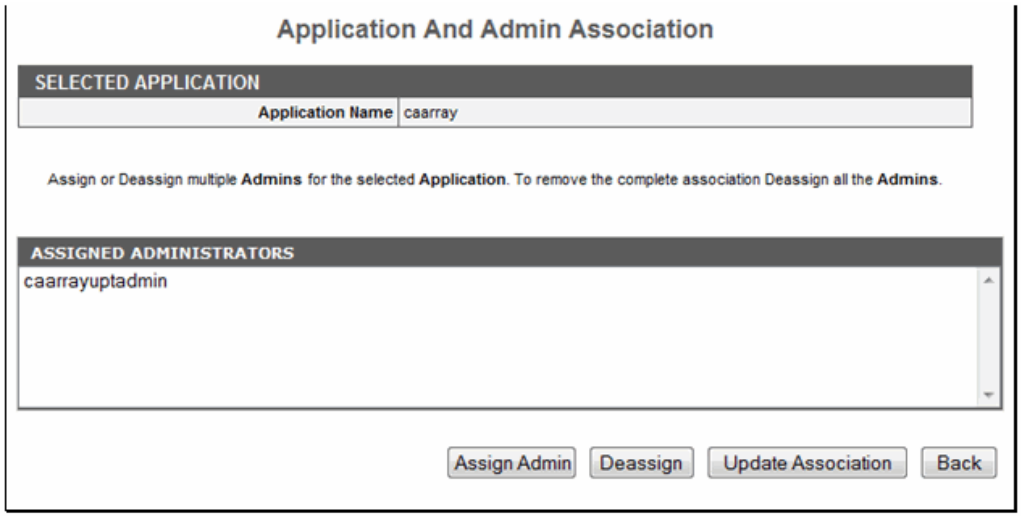


Figure 8.4 The user to whom you assign an admistrative role displays in the Assigned Administration section of the page

10. To actually associate the admin user with the caArray application **you must click the Update Associate button**. This step is crucial, and if you forget to do it you will need to go through the Assign Admin workflow again to make sure it sticks.

Figure 8.5 shows an example of a successfully assigned admin user.

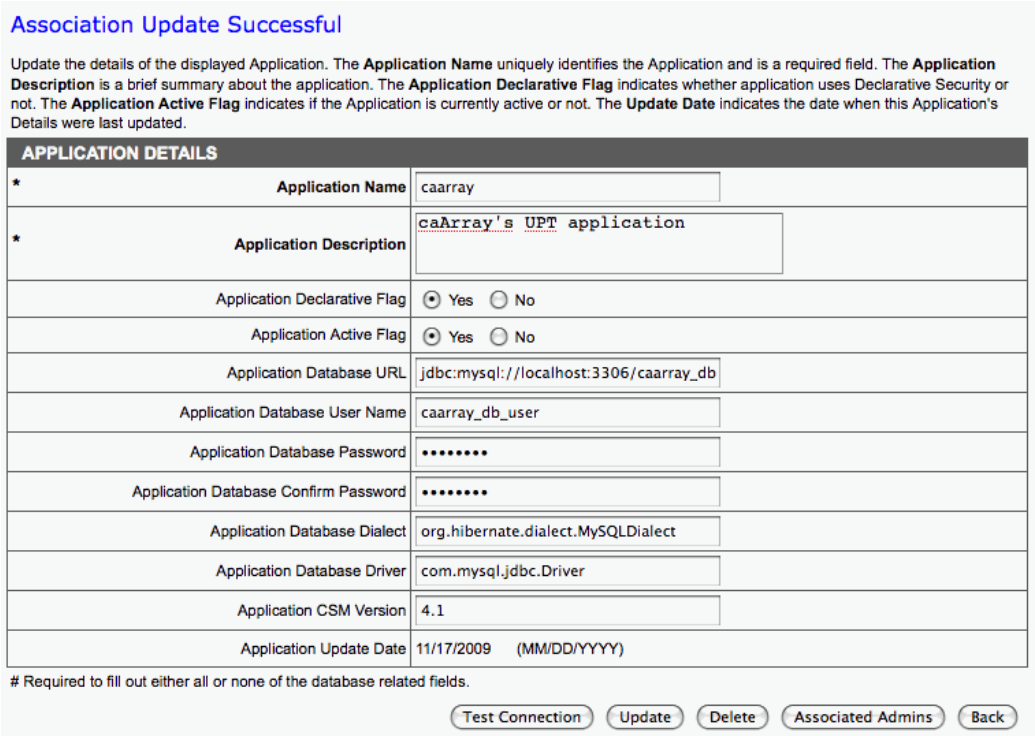


Figure 8.5 UPT form for entering caArray association information

At this point, you can log out of UPT, and log in again as the caArray system administrator..

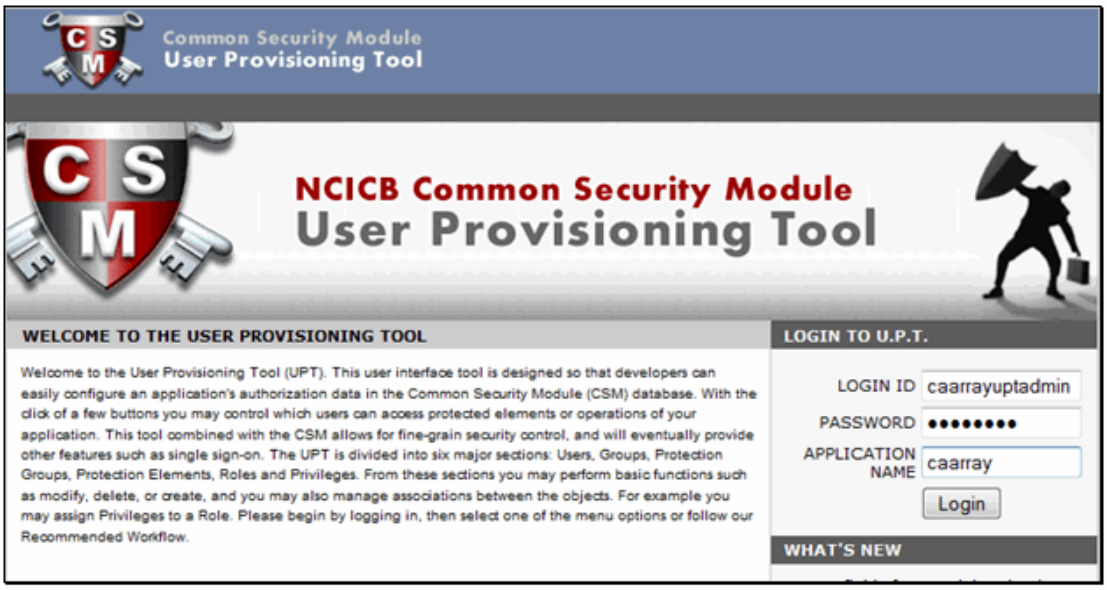


Figure 8.6 UPT login page

Upon successful login, UPT opens to the UPT Welcome page (*Figure 8.7*). If you cannot log in for some reason, login as `superadmin/changeme/csmupt`, and go back to the caArray application page in UPT to make sure that you created and associated the caArray system administrator correctly.

Anytime you want to perform user management tasks for the caArray application like creating users and assigning/changing user roles, you would log into UPT in this way.

Note that you **should not** use protection elements or protection groups directly in UPT, nor should you try to create caArray collaboration groups within UPT.

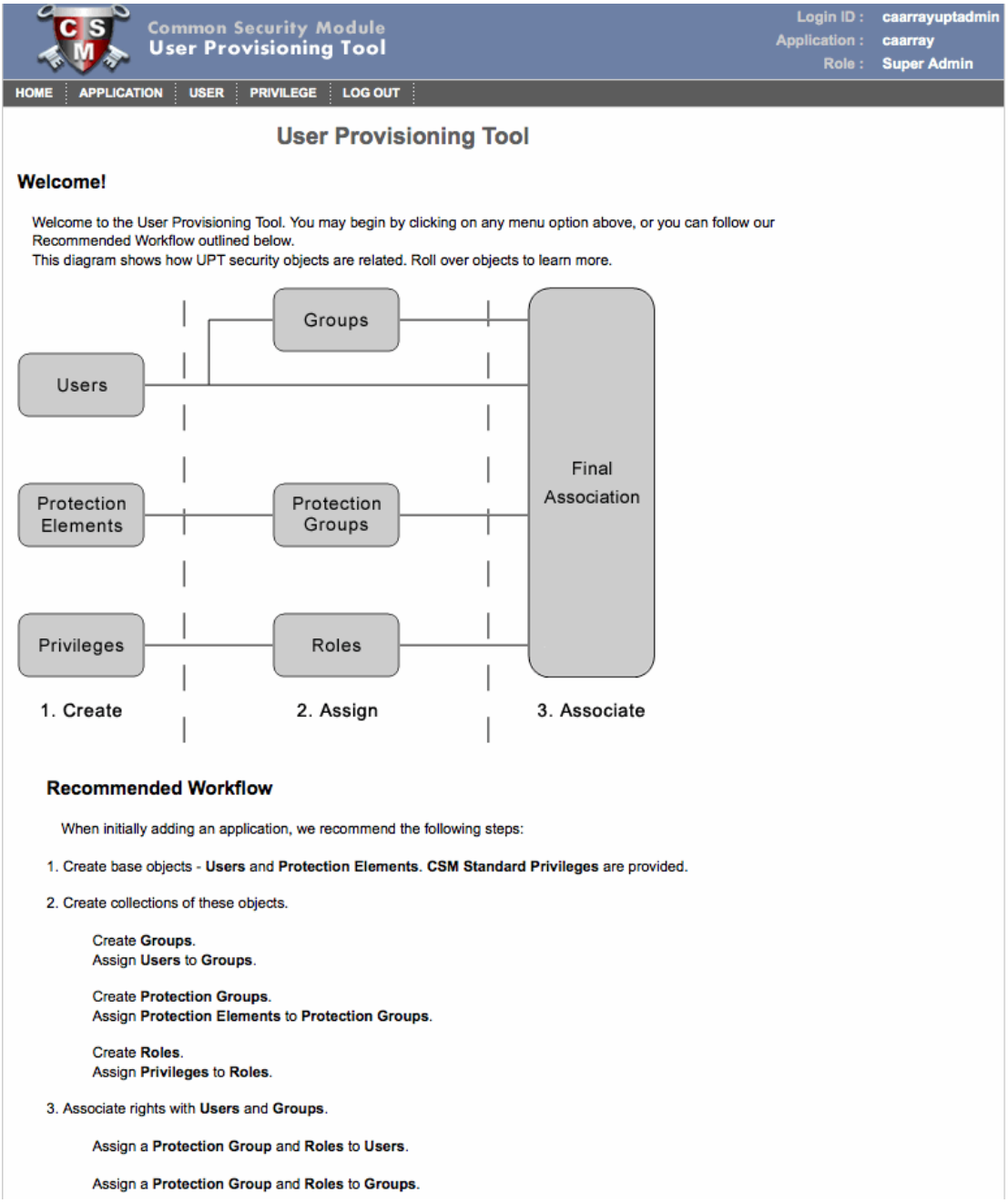


Figure 8.7 UPT Welcome page displaying UPT security diagram; visible only to a successfully logged in caArray UPT administrator

User management changes that are applied to the caArray application by UPT are effected directly to the caArray database, using the database connection information that is specified in the application configuration within UPT. For example, when you create a new caArray user within UPT, it inserts a new row in the csm_user table within the caArray database schema. User role changes are similarly performed via inserts, deletions, and updates within the `csm_*` caArray schema tables.

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 (*Table 8.1*). When the account is registered and roles are assigned, the user can fully access caArray according to the roles provisioned.

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. Can change ownership of experiments and/or collaboration groups.
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 2.3
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 2.3
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 2.3

Table 8.1 Roles within caArray and permissible tasks within those roles

Role	Description	Permissible Actions
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 *Extracting Data Programmatically by API* on page 94.

Table 8.1 Roles within caArray and permissible tasks within those roles (Continued)

Managing Experiment "Ownership" and Group Access

Any user with the System Administrator role can change ownership of experiments and/or collaboration groups. To do so, click the **Manage Ownership** link in the left sidebar. Search for a user of choice and review the experiment(s) and collaboration group(s) owned by that user. You can then select one or more of these assets and change their ownership to a different user.

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 8.8*).

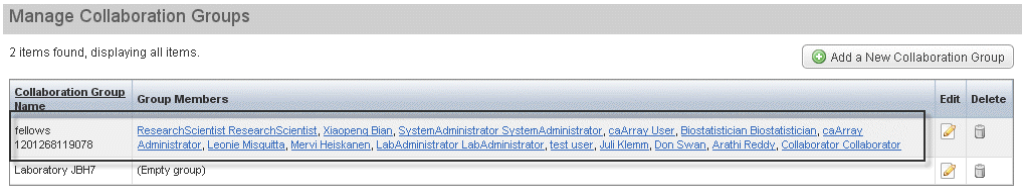


Figure 8.8 Collaboration groups are listed on the Manage Collaboration Groups page

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

   (Add a New Group Member).
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

   (Filter) 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.
8. 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.
9. 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 Collaboration 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:

1. From the Manage Collaboration Groups page, click the **Edit** icon ( ) corresponding to the group you select.
2. 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.

3. To perform these edits, follow the instructions in *Table 8.2*:.

Edit Function	*Description*
Edit the group name	Enter new name in the Group Name text box.
Add a new group member	**Note:** The new member must already have a valid caArray user account. 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.
Remove a group member	On the Collaboration Group page, in the **Remove** column, click the Remove 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.
Delete the group	To delete the entire Collaboration Group, click the **Delete** button (Delete) at the bottom center of the page.

Table 8.2 Collaboration group editing functions

Audit Log

If you log in with the system administrator permissions, caArray includes an audit trail feature that allows you to monitor experiment/sample permissions, experiment locking/unlocking and collaborator group changes.

To use the audit feature, follow these steps:

1. Log in to caArray as the system administrator.
2. On the left sidebar, click the **Audit Log** option under Management.

The Security Audit Log page that opens displays a log of all auditable actions available in the repository (*Figure 8.9*).

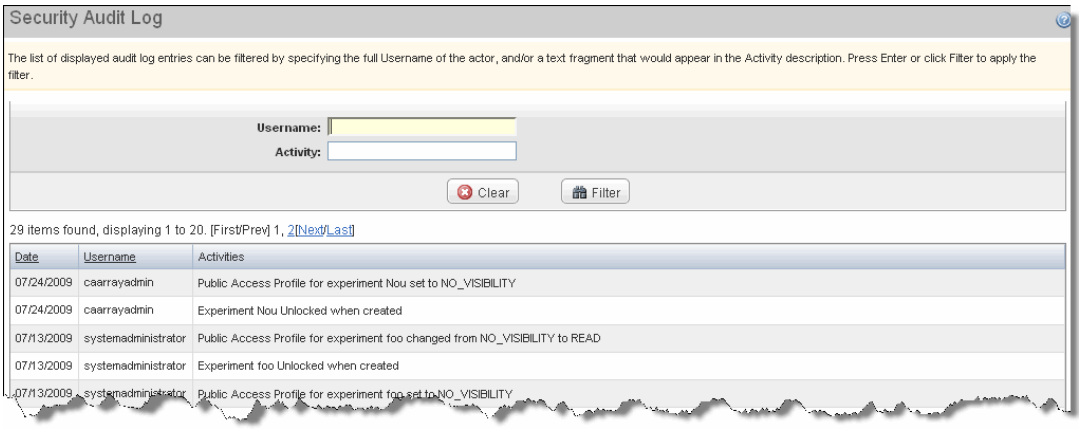


Figure 8.9 Audit log

Besides experiment/sample permissions changes, experiment locking/unlock and collaborator group changes, the information includes the date of the change, the user who made the change, and a description of the change (activity).

3. To filter the log view by user and/or by activity, enter the username and/or activity by which you want to filter the information in the text boxes provided. Click **Filter**.

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:

For information about importing MAGE-TAB documents into caArray, see *Importing MAGE-TAB Data* on page 87.

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 79 and *Importing MAGE-TAB Data* on page 87.

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 the previous section, 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 75), 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 2.4 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, usually indicated by 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 usually 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 *Term Source REF in an SDRF* on page 114.

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 *Term Source REF in an SDRF* on page 114.

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. 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.
3. 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.

4. 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
5. 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
6. 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. Illumina Data CSV files contain hybridization names that are implicit. The SDRF validation will fail if it does not contain these hybridization names.
2. Genepix GPR files must be accompanied by a MAGE-TAB SDRF and IDF; otherwise the import/validation will fail.
3. 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)

4. 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

5. 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

6. An Array Design Name column is unsupported. Use an Array Design REF instead.

7. 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.

8. 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
IMPORTING DATA FILES

caArray supports the storage of a wide variety of data files and enables the association of those data files to biomaterials (samples) and hybridizations. Depending on the type of data, there are specific rules that govern the association of data to biomaterials and hybridizations. In addition, caArray parses certain types of data files and stores discrete values in its repository, enabling external client applications to retrieve subsets of that data for further analysis. Upon import, unparsed file types end up in the "Imported, not Parsed" state, while parsed file types end up in the "Imported" state.

This appendix describes the rules governing validation and import of all the data file types supported by caArray. It is organized by provider (vendor or software).

Affymetrix

All Affymetrix data files can be imported with or without an accompanying MAGE-TAB IDF and SDRF. If accompanied by an IDF and SDRF, the data files are associated to biomaterials and hybridizations as specified in the SDRF. If no accompanying IDF-SDRF is present, then biomaterial-hybridization-data chains are generated according to one of the three user-selectable options as described in *Chapter 6 Submitting Data to an Experiment*.

Table B.1 summarizes the validation and import rules for Affymetrix data files.

Provider	File Format	MAGE-TAB required?	Raw or Derived?	Parsed?	Parsing Details
Affymetrix	DAT	no	raw	no	N/A
Affymetrix	EXP	no	derived	no	N/A
Affymetrix	TXT	no	derived	no	N/A
Affymetrix	RPT	no	derived	no	N/A
Affymetrix	CEL	no	raw	yes	GCOS binary/XDA and Command Console/AGCC/ Calvin fomats
Affymetrix	CHP (gene expression)	no	derived	yes	GCOS binary/XDA and Command Console/AGCC/ Calvin fomats;MAS5 and RMA/PLIER algorithms
Affymetrix	CHP (genotyping)	no	derived	yes	GCOS binary/XDA and Command Console/AGCC/ Calvin fomats; Birdseed, BRLMM and Axiom GT algorithms
Affymetrix	CNCHP	no	derived	yes	CN4 (only copy number) & CN5 (copy number and LOH) fomats

Table B.1 Affymetrix data file validation and import rules

Affymetrix files that caArray imports can be further described as follows.

File types imported without parsing:

- DAT — Data file containing raw image data (pixel intensity values).
- EXP — Data file containing information about experimental conditions and protocols.
- RPT — Report File summarizing data quality information generated from the CHP file.
- TXT - Exported file generated from the CHP file in text format.

As raw data, DAT files can appear only in the Array Data File column of an accompanying SDRF (if present). As derived data, the EXP, RPT and TXT files can appear only in the Derived Array Data File column of an accompanying SDRF (if present).

File types imported and parsed:

- CEL — Data file containing information about the intensity values of the individual probes.
- CHP — Data file containing summary information of the probe sets, including normalized intensity values.
- CNCHP — Level 2 derived data file containing copy number and (optionally) LOH data.

As raw data, CEL files can appear only in the Array Data File column of an accompanying SDRF (if present). As derived data, the CHP and CNCHP files can appear only in the Derived Array Data File column of an accompanying SDRF (if present).

GenePix

GenePix files are 2-or-more color experiments. An accompanying MAGE-TAB file set (IDF and SDRF) is required to ensure correct biomaterial-hybridization-data associations.

Table *B.2* summarizes the validation and import rules for Genepix data files.

Provider	***File Format***	***MAGE-TAB required?***	***Raw or Derived?***	***Parsed?***	***Parsing Details***
GenePix	GPR	yes	derived	yes	GPR file format version 3.0 is supported.

Table B.2 Genepix data file validation and import rules

There are no unparsed GenePix formats. caArray imports and parses the GenePix GPR (GenePix Results) file type version 3.0. GPR files are in Axon Text File (ATF) format, and contain general information about image acquisition and analysis, as well as the data extracted from each individual feature. As derived data, GPR files can appear only in the Derived Array Data File column of an accompanying SDRF (if present).

The file format is described in detail at this site:

http://www.moleculardevices.com/pages/software/gn_genepix_file_formats.html#gpr

Illumina

All Illumina data files can be imported with or without an accompanying MAGE-TAB IDF and SDRF.

Table *B.3* summarizes the validation and import rules for Illumina data files.

Provider	***File Format***	***MAGE-TAB required?***	***Raw or Derived?***	***Parsed?***	***Parsing Details***	***Biomaterials Data Association***
Illumina	DAT	no	raw	no	N/A	
Illumina	Raw TXT	no	raw	no	N/A	
Illumina	Derived TXT	no	derived	no	N/A	
Illumina	CSV	no	derived	yes	caArray parses several tab-delimited columns: Probe_ID/ID_REF is mandatory if the corresponding array design is in BGX/TXT format; otherwise, TargetID (mandatory) and ProbeId are expected. The following columns are present for each hybridization: AVG_Signal (mandatory), MIN_Signal, MAX_Signal, NARRAYS, ARRAY_STDEV, BEAD_STDEV, Avg_NBEADS, Detection/ DetectionPval/Detection Pval (mandatory). These columns will be either prefixed with "<hybridization_name>." or suffixed with "-<hybridization_name>". Multiple hybridizations are typically represented in a single file.	If the data files are imported without an accompanying MAGE-TAB file set, then biomaterial-hybridization chains will be auto-generated according to the hybridization names present in the data files themselves. If accompanied by MAGE-TAB, caArray checks hybridization names in the SDRF against hybridization names within the data files. If the SDRF is missing one or more of the hybridization names in the data files, validation/import will fail.

Table B.3 Illumina validation and data import rules

Provider	File Format	MAGE-TAB required?	Raw or Derived?	Parsed?	Parsing Details	Biomaterials Data Association
Illumina	Sample probe profile TXT	no	derived	yes	caArray parses several tab-delimited columns: Probe_ID/ID_REF is mandatory if the corresponding array design is in BGX/TXT format; otherwise, TargetID (mandatory) and ProbeId are expected. The following columns are present for each hybridization: AVG_Signal (mandatory), MIN_Signal, MAX_Signal, NARRAYS, ARRAY_STDEV, BEAD_STDEV, Avg_NBEADS, Detection/ DetectionPval/Detection Pval (mandatory). These columns will be either prefixed with "<hybridization_name>." or suffixed with "-<hybridization_name>". Multiple hybridizations are typically represented in a single file.	If the data files are imported without an accompanying MAGE-TAB file set, then biomaterial-hybridization chains will be auto-generated according to the hybridization names present in the data files themselves. If accompanied by MAGE-TAB, caArray checks hybridization names in the SDRF against hybridization names within the data files. If the SDRF is missing one or more of the hybridization names in the data files, validation/import will fail.
Illumina	Genotyping processed data matrix TXT	no	derived	yes	caArray parses the following tab-delimited columns:IlmnID/ID/ID_REF (mandatory), <hybridization_name> (mandatory), GC_SCORE, Theta, R, B_Allele_Freq, Log_R_Ratio. The last 5 quantitation types can optionally be prefixed with "<hybridization_name>.". Multiple hybridizations are typically represented in a single file.	If the data files are imported without an accompanying MAGE-TAB file set, then biomaterial-hybridization chains will be auto-generated according to the hybridization names present in the data files themselves. If accompanied by MAGE-TAB, caArray checks hybridization names in the SDRF against hybridization names within the data files. If the SDRF is missing one or more of the hybridization names in the data files, validation/import will fail.

Table B.3 Illumina validation and data import rules

Illumina files that caArray imports can be further described as follows.

File types imported without parsing:

- IDAT — Data file containing raw image data.
- Raw TXT - Plain text file containing raw signal data in one of a variety of formats.
- Derived TXT - Plain text file containing processed signal data in one of a variety of formats that don't conform to any of the parsed file types described later in this section.

As raw data, IDAT and Raw TXT files can appear only in the Array Data File column of an accompanying SDRF (if present). As derived data, Derived TXT files can appear only in the Derived Array Data File column of an accompanying SDRF (if present).

File types imported and parsed:

- CSV (gene expression) - a comma-separated-values file containing processed signal values. This was generated by older versions of BeadStudio (pre 3.0).
- Sample probe profile TXT (gene expression) - a tab-separated values file containing processed signal values. This is generated by newer versions of BeadStudio and GenomeStudio.
- Genotyping processed data matrix TXT (genotyping) - a tab-separated values file containing log ratio values.

As derived data, all three of the above file types can appear only in the Derived Array Data File column of an accompanying SDRF (if present). The following sub-sections describe these file formats and their parsing in detail.

See the example Illumina files in the following topics:

- *CSV (Gene Expression)* on page 130
- *Sample Probe Profile TXT (Gene Expression)* on page 131
- *Genotyping Processed Data Matrix TXT* on page 132

CSV (Gene Expression)

caArray parses several comma-delimited columns from Illumina gene expression CSV files. TargetID (mandatory) and ProbeId identify the probes. The columns described in Table *B.3* are present for each hybridization. These columns will be suffixed with "-<hybridization_name>". Multiple hybridizations are typically represented in a single file. A snippet from an example file is shown below:

Illumina Inc. BeadStudio version 1.5.1.3,,

Normalization = cubic spline,,

Array Content = Human_WG-6.xml,,

Error Model = none,,

DateTime = 9/30/2005 6:39 PM,,

Local Settings = en-US,,,

TargetID,AVG_Signal-A1,BEAD_STDEV-A1,Avg_NBEADS-A1,Detection-A1,AVG_Signal-B1,BEAD_STDEV-B1,Avg_NBEADS-B1,Detection-B1,AVG_Signal-C1,BEAD_STDEV-C1,Avg_NBEADS-C1,Detection-C1,AVG_Signal-D1,BEAD_STDEV-D1,Avg_NBEADS-D1,Detection-D1,AVG_Signal-A2,BEAD_STDEV-A2,Avg_NBEADS-A2,Detection-A2,AVG_Signal-B2,BEAD_STDEV-B2,Avg_NBEADS-B2,Detection-B2,AVG_Signal-C2,BEAD_STDEV-C2,Avg_NBEADS-C2,Detection-C2,AVG_Signal-D2,BEAD_STDEV-D2,Avg_NBEADS-D2,Detection-D2,AVG_Signal-A3,BEAD_STDEV-A3,Avg_NBEADS-A3,Detection-A3,AVG_Signal-B3,BEAD_STDEV-B3,Avg_NBEADS-B3,Detection-B3,AVG_Signal-D3,BEAD_STDEV-D3,Avg_NBEADS-D3,Detection-D3,AVG_Signal-A4,BEAD_STDEV-A4,Avg_NBEADS-A4,Detection-A4,AVG_Signal-B4,BEAD_STDEV-B4,Avg_NBEADS-B4,Detection-B4,AVG_Signal-C4,BEAD_STDEV-C4,Avg_NBEADS-C4,Detection-C4,AVG_Signal-D4,BEAD_STDEV-D4,Avg_NBEADS-D4,Detection-D4,AVG_Signal-A5,BEAD_STDEV-A5,Avg_NBEADS-A5,Detection-A5,AVG_Signal-B5,BEAD_STDEV-B5,Avg_NBEADS-B5,Detection-B5,AVG_Signal-C5,BEAD_STDEV-C5,Avg_NBEADS-C5,Detection-C5,AVG_Signal-D5,BEAD_STDEV-D5,Avg_NBEADS-D5,Detection-D5

GI_10047089-S,5.8,3.3,45,0.84179301,105.1,6.7,48,1,14.7,3.5,56,0.98681608,56.7,4,50,0.9993408,0.4,3,39,0.53460778,124.8,6.5,39,1,18.5,4.1,42,0.99406724,64.5,5.8,37,1,4.1,3.3,48,0.77982861,91.7,5.9,40,1,70.1,5.2,39,0.9993408,2.8,3.8,36,0.70731707,102,6.5,44,1,16.7,2.7,40,0.99406724,57.4,3.7,47,1,7,4,46,0.90112063,105.6,6.7,31,1,15.1,5.2,35,0.98879367,68.7,5.4,34,1

Sample Probe Profile TXT (Gene Expression)

caArray parses several tab-delimited columns, described in Table *B.3*, from Illumina sample prove profile gene expression TXT files. These columns will be either prefixed with "<hybridization_name>." or suffixed with "-<hybridization_name>". Multiple hybridizations are typically represented in a single file. A snippet from an example file is shown below:

Probe_Id	Patient 1-Tumor.AVG_Signal	Patient 1-Tumor.NARRAYS Patient 1-Tumor.ARRAY_STDEV	Patient 1-Tumor.Detection Pval	Patient 1-Tumor.BEAD_STDERR	Patient 1-Tumor.Avg_NBEADS	Patient 2-Tumor.AVG_Signal	Patient 2-Tumor.NARRAYS Patient 2-Tumor.ARRAY_STDEV	Patient 2-Tumor.Detection Pval	Patient 2-Tumor.BEAD_STDERR	Patient 2-Tumor.Avg_NBEADS

ILMN_2136495	111.1856	1	NaN	0	6.041573	36	87.39001	1	NaN	0	6.327084	33

A snippet from another example file demonstrates the second variant:

TargetID	ProbeID MIN_Signal-1745080067_B AVG_Signal-1745080067_B MAX_Signal-1745080067_B NARRAYS-1745080067_B	ARRAY_STDEV-1745080067_B BEAD_STDEV-1745080067_B Avg_NBEADS-1745080067_B Detection-1745080067_B MIN_Signal-1745080067_A AVG_Signal-1745080067_A MAX_Signal-1745080067_A NARRAYS-1745080067_A	ARRAY_STDEV-1745080067_A	BEAD_STDEV-1745080067_A Avg_NBEADS-1745080067_A Detection-1745080067_A

0610005I04 580022 164.9 164.9 164.9 1 NaN 61.030 33 0.51703 164.9 164.9 164.9 1 NaN 57.871 29 0.73723

Genotyping Processed Data Matrix TXT

caArray parses the genotyping processed data matrix tab-delimited columns described in Table *B.3*, from Illumina processed data matrix genotyping TXT files.

Multiple hybridizations are typically represented in a single file. A snippet from an example file is shown below:

IlmnID HYB1 GC_SCORE Theta R B_Allele_Freq Log_R_Ratio HYB2 GC_SCORE Theta R B_Allele_Freq Log_R_Ratio

MitoA10045G-13273284_B_R_IFB1141652022:0 AA 0.3197343 0.006295365 6.168365 0.000579326 0.4340365 AA 0.3197343 0.005859224 4.608023 9.934365E-05 0.01329622

A snippet from another example file demonstrates the second variant:

ID TCGA-06-0119-01A-08D TCGA-06-0119-01A-08D.GC_SCORE TCGA-06-0119-01A-08D.Theta TCGA-06-0119-01A-08D.R TCGA-06-0119-01A-08D.B_Allele_Freq TCGA-06-0119-01A-08D.Log_R_Ratio TCGA-06-0119-01A-08D TCGA-06-0119-01A-08D.GC_SCORE TCGA-06-0119-01A-08D.Theta TCGA-06-0119-01A-08D.R TCGA-06-0119-01A-08D.B_Allele_Freq TCGA-06-0119-01A-08D.Log_R_Ratio

MitoA10045G-13273284_B_R_IFB1141652022:0 AA 0.3197343 0.006295365 6.168365 0.000579326 0.4340365 AA 0.3197343 0.005859224 4.608023 9.934365E-05 0.01329622

Agilent

Agilent raw TXT files from aCGH and 2-color gene expression assays must be accompanied by a MAGE-TAB IDF and SDRF to ensure correct biomaterial-hybridization-data associations. All other Agilent data files can be imported with or without an accompanying MAGE-TAB IDF and SDRF.

Table *B.4* summarizes the validation and import rules for Agilent data files.

Provider	*File Format*	*MAGE-TAB required?*	*Raw or Derived?*	*Parsed?*	*Parsing Details*
Agilent	TSV	no	raw	no	N/A
Agilent	Derived TXT	no	derived	no	N/A

Table B.4 Agilent validation and import data import rules

Provider	File Format	MAGE-TAB required?	Raw or Derived?	Parsed?	Parsing Details
Agilent	Raw TXT (aCGH and 2-color gene expression)	yes	raw	yes	caArray parses the FEATURES section of this file, and recognizes the following tab-delimited columns: ProbeName, LogRatio, LogRatioError, PValueLogRatio, gProcessedSignal, rProcessedSignal, gProcessedSigError, rProcessedSigError, gMedianSignal, rMedianSignal.
Agilent	Raw TXT (1-color gene expression)	no	raw	yes	caArray parses the FEATURES section of this file, and recognizes the following tab-delimited columns: ProbeName, gProcessedSignal, gProcessedSigError, gMedianSignal.
Agilent	Raw TXT (miRNA)	no	raw	yes	caArray parses the FEATURES section of this file, and recognizes the following tab-delimited columns: ProbeName, gTotalProbeSignal, gTotalProbeError, gTotalGeneSignal, gTotalGeneError, gIsGeneDetected.

Table B.4 Agilent validation and import data import rules

Agilent files that caArray imports can be further described as follows.

File types imported without parsing:

- TSV
- Derived TXT - Plain text file containing processed data in one of a variety of formats that don't conform to any of the parsed file types described later in this section.

As raw data, TSV files can appear only in the Array Data File column of an accompanying SDRF (if present). The Derived TXT files can appear only in the Derived Array Data File column of an accompanying SDRF (if present).

File types imported and parsed:

- Raw TXT (aCGH and 2-color gene expression) - a tab-separated-values text file containing signal data generated by the Agilent Feature Extraction Software.
- Raw TXT (single color gene expression) - a tab-separated-values text file containing signal data generated by the Agilent Feature Extraction Software.
- Raw TXT (miRNA) - a tab-separated-values text file containing signal data generated by the Agilent Feature Extraction Software.

As raw data, all three of the above file types can appear only in the Array Data File column of an accompanying SDRF (if present).

Nimblegen

All Nimblegen data files can be imported with or without an accompanying MAGE-TAB IDF and SDRF files. If accompanied by an IDF and SDRF, the data files are associated to biomaterials and hybridizations as specified in the SDRF. If no accompanying IDF-SDRF is present, then biomaterial-hybridization-data chains are generated as described in *Auto-Generated Missing Biomaterials in MAGE-TAB* on page 112 and *Protocols Associated Intelligently* on page 112.

Table *B.5* summarizes the validation and import rules for Nimblegen data files.

Provider	***File Format***	***MAGE-TAB required?***	***Raw or Derived?***	***Parsed?***	***Parsing Details***
Nimblegen	GFF	no	derived	no	N/A
Nimblegen	Raw TXT	no	raw	no	N/A
Nimblegen	Derived TXT	no	derived	no	N/A
Nimblegen	Pair Report TXT	no	raw	yes	caArray recognizes the following tab-delimited columns: GENE_EXPR_OPTION, SEQ_ID, PROBE_ID, MATCH_INDEX, PM, MM, X, Y. "<GENE_EXPR_OPTION>|<SEQ_ID>|<PROBE_ID>" is used to match to the PhysicalProbe name from the NDF array design.

Table B.5 Nimblegen data file validation and import rules

Provider	File Format	MAGE-TAB required?	Raw or Derived?	Parsed?	Parsing Details
Nimblegen	Normalized Pair Report TXT	no	derived	yes	caArray recognizes the following tab-delimited columns: GENE_EXPR_OPTION, SEQ_ID, PROBE_ID, MATCH_INDEX, PM, MM, X, Y. "<GENE_EXPR_OPTION>|<SEQ_ID>|<PROBE_ID>" is used to match to the PhysicalProbe name from the NDF array design.

Table B.5 Nimblegen data file validation and import rules

Nimblegen files that caArray imports can be further described as follows.

File types imported without parsing:

- GFF
- Raw TXT - Plain text file containing raw data in one of a variety of formats that don't conform to any of the parsed file types described later in this section.
- Derived TXT - Plain text file containing derived/processed data in one of a variety of formats that don't conform to any of the parsed file types described later in this section.

As raw data, Raw TXT files can appear in only the Array Data File column of an accompanying SDRF (if present). As derived data, the GFF and Derived TXT files can appear in only the Derived Array Data File column of an accompanying SDRF (if present).

File types imported and parsed:

- Pair Report TXT - a tab-separated-values text file containing signal intensity information for a single channel.
- Normalized Pair Report TXT - a tab-separated-values text file containing normalized signal intensities output by the RMA algorithm.

As raw data, Pair Report TXT files can appear in only the Array Data File column of an accompanying SDRF (if present). As derived data, Normalized Pair Report TXT files can appear in only the Derived Array Data File column of an accompanying SDRF (if present)).

Miscellaneous Providers

caArray validates and imports several other file types as long as they comply with a caArray-prescribed format.

Copy Number MAGE-TAB Data Matrix TXT

This is a special MAGE-TAB Data Matrix file that has copy number data in a caArray-prescribed format, and an accompanying MAGE-TAB file set (IDF and SDRF) is required. caArray expects the first header row to contain tab-delimited Hybridization REFs that refer to Hybridization Names in the corresponding SDRF. The rest of the file is expected to contain the following tab-delimited columns: Reporter REF (probe name), Chromosome, Position, and Log2Ratio.

A snippet from an example file is shown in *Figure B.1*:

Hybridization REF			my_hybridization_1	my_hybridization_2
Reporter REF	Chromosome	Position	Log2Ratio	Log2Ratio
A_16_P37638626	6	62964904	-0.447480618	-0.457480618
A_18_P10656728	10	19811167	-0.258419751	-0.268419751
A_18_P23155090	3	127363084	-0.303141107	-0.313141107
A_18_P14566787	3	195691534	-0.306512915	-0.316512915
A_14_P128705	5	77351300	-0.160818764	-0.170818764
A_16_P37904891	7	2275926	-0.269197169	-0.279197169
A_16_P16670574	4	45262920	-0.320056597	-0.330056597

Figure B.1 Example MAGE-TAB copy number IDF file

Other Providers

Table *B.6* summarizes the validation and import rules for all other data files supported by caArray.

Provider	***File Format***	***MAGE-TAB required?***	***Raw or Derived?***	***Parsed?***
ImaGene	TIF	no	raw	no
ImaGene	TXT	no	derived	no
GEO	SOFT	no	raw	no
GEO	GSM	no	raw	no
ScanArray	CSV	no	raw	no

Table B.6 Miscellaneous data file validation and import rules

caArray also imports without parsing MAGE-TAB Data Matrix (not the Copy Number Data Matrix format described earlier).

As raw data, ImaGene TIF, GEO SOFT, GEO GSM and ScanArray CSV files can appear only in the Array Data File column of an accompanying SDRF (if present). As derived data, ImaGene TXT files can appear only in the Derived Array Data File column of an accompanying SDRF (if present). MAGE-TAB Data Matrix files can represent data from any provider, and can appear in either the Array Data Matrix File (raw) or Derived Array Data Matrix File (derived) column of an accompanying SDRF.

APPENDIX C

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

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V