



caArray Cancer Array Informatics

Software Overview

NCI Center for Bioinformatics

May 17, 2004

Agenda

- Project goals
- Current project status
- Experiment submission/wire frames
- caBIG compliance
- Architecture & design overview
- Q&A

caArray Goals

- Simplified data submission and annotation
 - Allow MAGE-ML import
 - Allow Affymetrix and GenePix native file uploads for Hybridizations
 - Provide MIAME 1.1 level annotations with intuitive interface
 - Controlled vocabularies (MGED ontology)
- Interoperability
 - MAGE-ML import and export
 - Open API
- Leverage existing open source technologies
 - Utilizes the MAGE-stk Java toolkit
 - Uses Apache/Jakarta Object Relational Bridge to map MAGE objects to relational database

Key Features (Phase I)

- ❑ MIAME 1.1 compliant
- ❑ MAGE-ML import and export
- ❑ Submission and retrieval of native Affymetrix and GenePix files
- ❑ Use of controlled vocabularies
- ❑ N-tier architecture
- ❑ Access via MAGE-OM API
- ❑ User-access management via the NCICB common security model

Key Features (Phase II)

- Usability enhancements based on user feedback
- Support for UCSF SPOT data
- User interface for User Management

Data Management

Gene Expression Data Portal (NCICB)

- Based on MIAME 1.0
- Public data repository
- Affymetrix, Spotted Array, CGH, and SAGE data

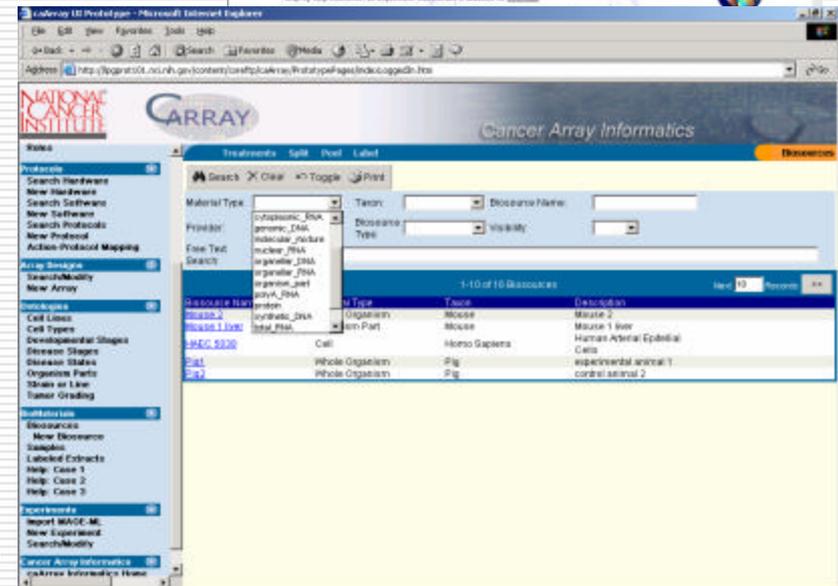
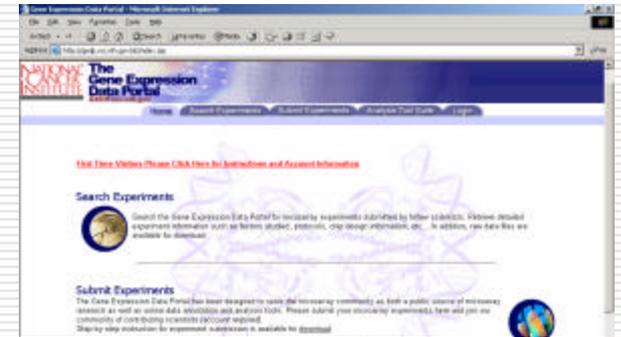
Available in September 2004:

caArray database (NCICB)

- Based on MIAME 1.1
- Public data repository
- Local deployment at cancer centers

MIAMExpress (EBI)

- MAGE-ML import to caArray database



Survey of Existing MicroArray Systems

- 19 Existing MicroArray systems compared on the following criteria
 - Computationally MIAME 1.1 compliant
 - MAGE-ML Import
 - MAGE-ML Export
 - Affymetrix native file upload support
 - Genepix native file upload support
 - Security and data sharing capabilities
 - Ability to modify data after submission
 - Add/Remove Hybridizations
 - Correct mistakes in annotations
 - Architecture/Open source
 - **Ease of integration with caCORE technologies**
- No existing system was found to meet these needs

Current Status

❑ Completed Analysis

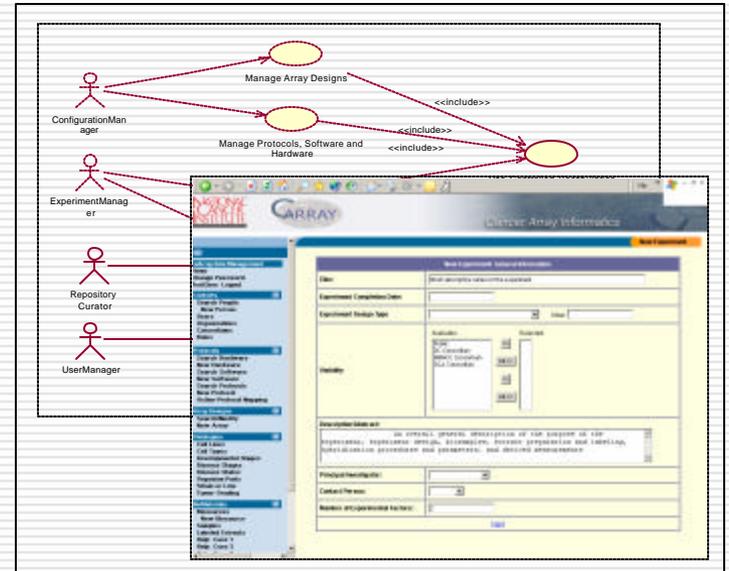
- Use case diagrams
- Wire frames

❑ Completed Design

- Sequence diagrams
- Class diagrams
- Reference

implementation /
proof of concept of protocol management

❑ Completed MAGE-stk persistence



caArray UI prototype

<http://caarray.nci.nih.gov/caARRAY/devdoc/caarraydbdocs>





- caArray Data Management
 - Home
 - Change Password
 - TestUser: Logout
- Contacts
 - Search People
 - New Person
 - Users
 - Organizations
 - Consortiums
 - Roles
- Protocols
 - Search Hardware
 - New Hardware
 - Search Software
 - New Software
 - Search Protocols
 - New Protocol
 - Action-Protocol Mapping
- Array Designs
 - Search/Modify
 - New Array
- Ontologies
 - Cell Lines
 - Cell Types
 - Developmental Stages
 - Disease Stages
 - Disease States
 - Organism Parts
 - Strain or Line
 - Tumor Grading

Welcome TestUser.

You currently have access to Public, SPORES, and DC data



caArray Data Management

Home
Change Password
TestUser: Logout

Contacts

Protocols

Search Hardware
New Hardware
Search Software
New Software
Search Protocols
New Protocol
Action-Protocol Mapping

Array Designs

BioMaterials

Experiments

Cancer Array Informatics

Search Clear Toggle Print

Hardware Type: Hardware Model: Visibility:

Hardware Make: Hardware Manufacture:

Free Text Search:

1-10 of 16 Hardwares

Next Records >>

Hardware Model	Hardware Type	Hardware Make	Hardware Manufacture
Hardware Model1	Array Scanner	Hardware Make 1	Hardware Manufacture 1
Hardware Model2	Array Scanner	Hardware Make 2	Hardware Manufacture 2
Fluidics Station 450	hybridization_station	Affymetrix	Affymetrix, Inc.
Standard HybChamber	hybridization_chamber	GeneMachines	GeneMachines
Themomix 1480	waterbath	B. Braun	Westshore Technologies, Inc.
GenePix® 4000B	array_scanner	Axon	Axon Instruments, Inc
Affymetrix 418 scanner	array_scanner	Affymetrix	Affymetrix, Inc.

Protocol Management

Dynamic Ontologies

The screenshot displays the 'caArray UI Prototype' in Microsoft Internet Explorer. The main page features the National Cancer Institute logo and 'CARRAY Cancer Array Informatics' branding. A search bar is visible with the text 'Search Cell Lines'. Below the search bar, there are input fields for 'Cell Line Name', 'Public Database (for reference)', 'Visibility', and 'Cell Line Accession Number'. A table lists search results, with one entry highlighted: '2E10-H2' from the 'ACC 178' database in the 'DSMZ' accession number. A 'Query Results' window is open, showing options to apply filters to the results and displaying detailed information for 'DSMZ_MUTZ:ACC 178'. The detailed information includes the accession number (ACC 178), cell line name (2E10-H2), a brief description, and a full description.

caArray Data Management

- Home
- Change Password
- TestUser: Logout
- Contacts
- Protocols
- Array Designs
- Search/Modify
- New Array
- Ontologies
- Cell Lines
- Cell Types
- Developmental Stages
- Disease Stages
- Disease States
- Organism Parts
- Strain or Line
- Tumor Grading
- BioMaterials
- Experiments
- Cancer Array Informatics
- caArray Informatics Home

Search Cell Lines

Search Clear

Cell Line Name: Visibility:

Public Database (for reference): Cell Line Accession Number:

1-10 of 16 Cell Lines Next Records >>

Cell Line Name	Database	Accession Number	Link
2E10-H2	ACC 178	DSMZ	Link to ACC 178 in DSMZ database

Query Results - Microsoft Internet Explorer

Address: [http://www.cabri.org/CABRI/srs-bin/wgetz?-newId+-e+-page+qResult+\[DSMZ_MUTZ-id:'ACC%20178'\]](http://www.cabri.org/CABRI/srs-bin/wgetz?-newId+-e+-page+qResult+[DSMZ_MUTZ-id:'ACC%20178'])

DSMZ_MUTZ:ACC 178

Apply Options to:

- selected results only
- unselected results only

Result Options

Link to related information:

Save results:

Display Options

CABRI:DSMZ_MUTZ

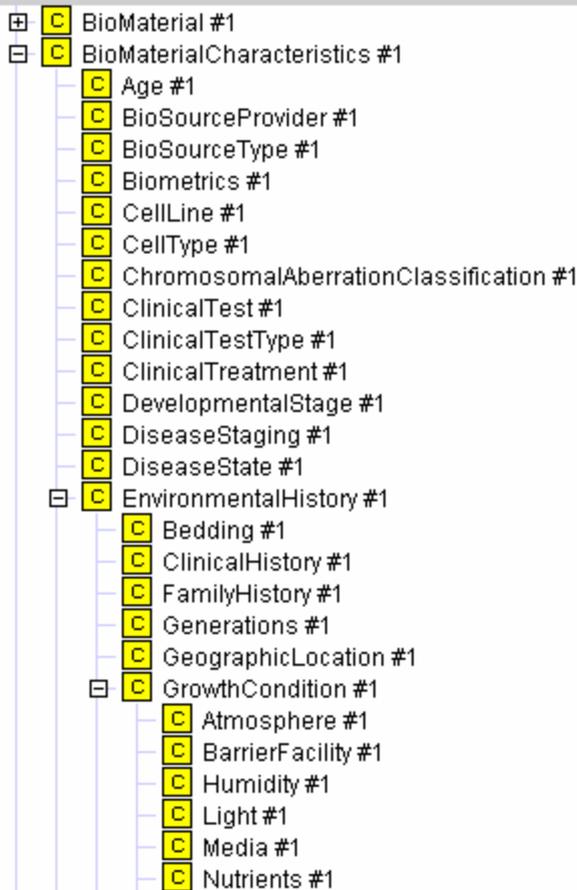
Accession_number ACC 178

Cell_line_name 2E10-H2

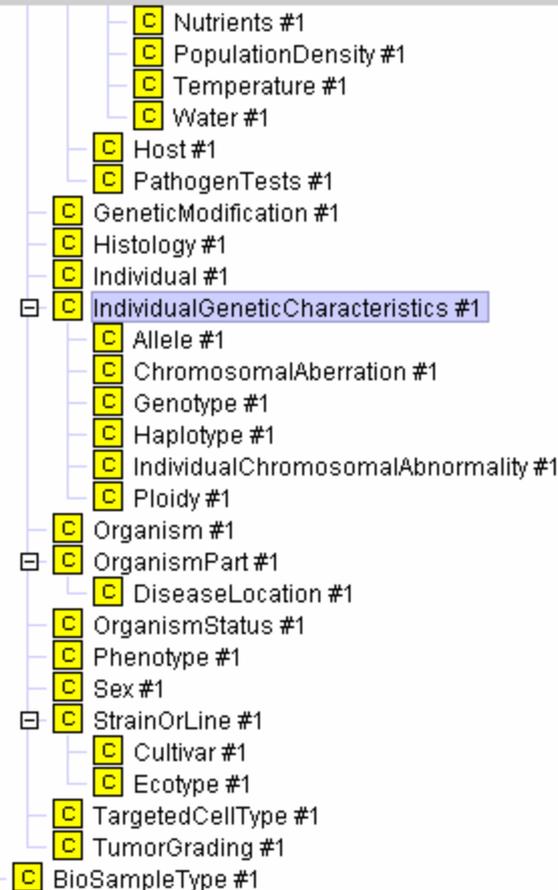
Brief_description confirmed as mouse with IEF of AST, NP established by fusion of SP2/0-AG14 myeloma cells with BALB/c mouse spleen cells immunized with purified canine adenovirus type 1 (CAV-1); the secreted antibody binds to and neutralizes CAV-1 and is type-specific

Description mouse hybridoma (anti-canine adenovirus type 1) established by fusion of SP2/0-AG14 myeloma cells with BALB/c mouse spleen cells immunized with purified canine adenovirus type 1 (CAV-1); the secreted antibody binds to and neutralizes CAV-1 and is type-specific

Hierarchy



Hierarchy



Supers

[-] BioMaterialCharacteristics #1

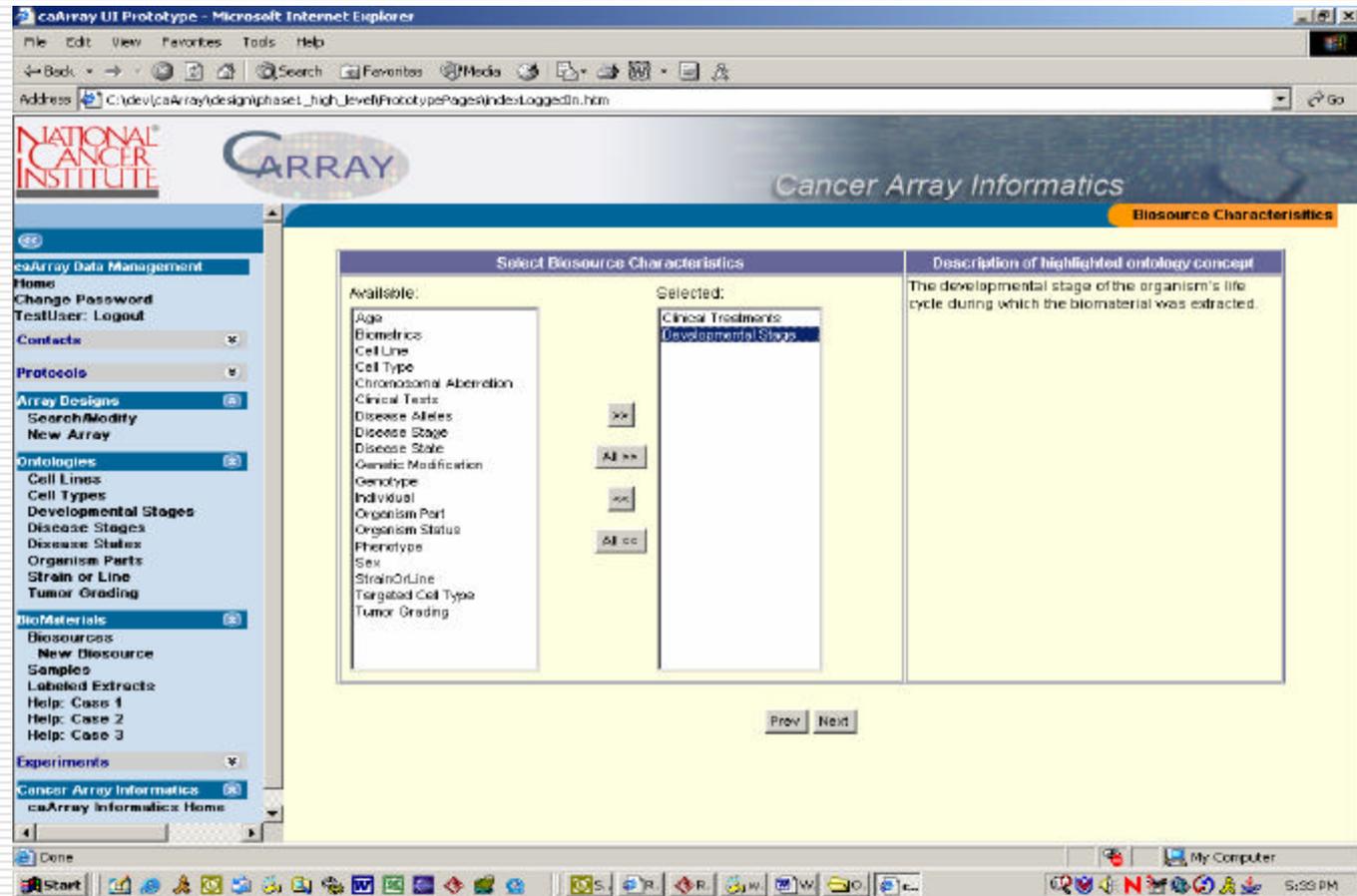
Supers

[-] BioMaterialCharacteristics #1

BioMaterialCharacteristics

BioMaterial Characteristics

Specify relevant characteristics



caArray UI Prototype - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address C:\dev\caarray\design\phase1_high_level\PrototypePages\index\loggedIn.htm

NATIONAL CANCER INSTITUTE **CARRAY** *Cancer Array Informatics*

Biosource: Mouse 2

Transfer Ownership **Duplicate** Delete Done

Biosource Information	
Biosource Name	Mouse 2
Biosource Description	Some description
Material Type	whole_organism
Organism	Mouse
Biosource Provider	Joe Doe
Biosource Type	paraffin_sample

Visibility: Public

Biosource Characteristics	
Age:	<input type="text"/> years Min Age: <input type="text"/> years Max Age: <input type="text"/> years Initial Time Point: <input type="text"/>
Biometrics:	Weight <input type="text"/> lbs Height <input type="text"/> cm Other: <input type="text"/>
Cell Line	<input type="text"/>
Cell Type	<input type="text"/>
Chromosomal	Type(s): <input type="text"/>

Cancer Array Informatics

Disease State	<input type="text"/>
Genetic Modification	Type: <input type="text"/> Description: <input type="text"/>
Genotype	<input type="text"/>
Histology	<input type="text"/>
Individual ID	<input type="text"/>
Organism Part	<input type="text"/>
Organism Status	<input type="text"/>
Phenotype	<input type="text"/>
Ploidy	<input type="text"/>
Sex	<input type="text"/>
Strain or Line	<input type="text"/>
Targeted Cell Type	<input type="text"/>
Tumor Grading	<input type="text"/>
Additional Description	<input type="text"/>

Save Prev Cancel

Done

Start

Resource Logout

Contacts

Protocols

Array Designs

Ontologies

BiMaterials

Biosources

Samples

Labeled Extracts

Help: Case 1

Help: Case 2

Help: Case 3

Experiments

Cancer Array Informatics

caArray Informatics Home

Done

Start

5:36 PM

5:37 PM



caArray Data Management

Home

Change Password

TestUser: Logout

Contacts

Protocols

Array Designs

BioMaterials

Biosources

New Biosource

Samples

Labeled Extracts

Help: Case 1

Help: Case 2

Help: Case 3

Experiments

Cancer Array Informatics

Biomaterial Treatment (biosample creation).

Biomaterial Information																			
Biomaterial creation description:	<input type="text" value="Mouse 2 Pancreas RNA extraction"/>																		
Select source biomaterial	<table border="0"> <tr> <td>Available:</td> <td></td> <td>Selected:</td> </tr> <tr> <td>HAEC 5038</td> <td>>></td> <td>Mouse 1 liver</td> </tr> <tr> <td>Pig 1</td> <td></td> <td>Mouse 2</td> </tr> <tr> <td>Pig 2</td> <td>All >></td> <td></td> </tr> <tr> <td></td> <td><<</td> <td></td> </tr> <tr> <td></td> <td>All <<</td> <td></td> </tr> </table>	Available:		Selected:	HAEC 5038	>>	Mouse 1 liver	Pig 1		Mouse 2	Pig 2	All >>			<<			All <<	
Available:		Selected:																	
HAEC 5038	>>	Mouse 1 liver																	
Pig 1		Mouse 2																	
Pig 2	All >>																		
	<<																		
	All <<																		
Name resulting biomaterial	<input type="text" value="Mouse 2 Pancreas RNA"/>																		
Type of resulting biomaterial	<input type="text" value="total_RNA"/>																		
Description of resulting biomaterial	<input type="text" value="Pancreas dissected from Mouse 2; total RNA extracted from pancreas"/>																		

Next

Cancel

New Experiment: (= study) General Information and Experimental Factors

caArray UI Prototype - Microsoft Internet Explorer

Address: C:\dev\caArray\design\phase1_high_level\PrototypePages\index.htm

NATIONAL CANCER INSTITUTE **CARRAY** *Cancer Array Informatics*

General Experiment Information

Title:	Experiment1	Modify
Experiment Date:	Date of completion of the experiment	
Experiment Type:	Normal vs. Diseased Comparison	
Visibility:	Public	
Description/Abstract:	An overall general description of the purpose of the experiment, experiment design, biosamples, extract preparation and labeling, hybridization procedures and parameters, and derived measurements	

Contacts

Principal Investigator:	Scott Melby	Modify
Contact:	Juergen Lorenz	

Experimental Factors/Variables

Factor	Factor Type	Scale	Levels	Actions
Age	Type 1	ordinal	1-12 Months 13-24 Months <input type="text"/>	Remove
Cell type	Type 2	nominal	Cell type1 Cell type2 <input type="text"/>	

[Add A New Factor](#)

File Upload

caArray UI Prototype - Microsoft Internet Explorer

Address: C:\dev\caArray\design\phase1_high_level\PrototypePages\index.htm

NATIONAL CANCER INSTITUTE **CARRAY** *Cancer Array Informatics*

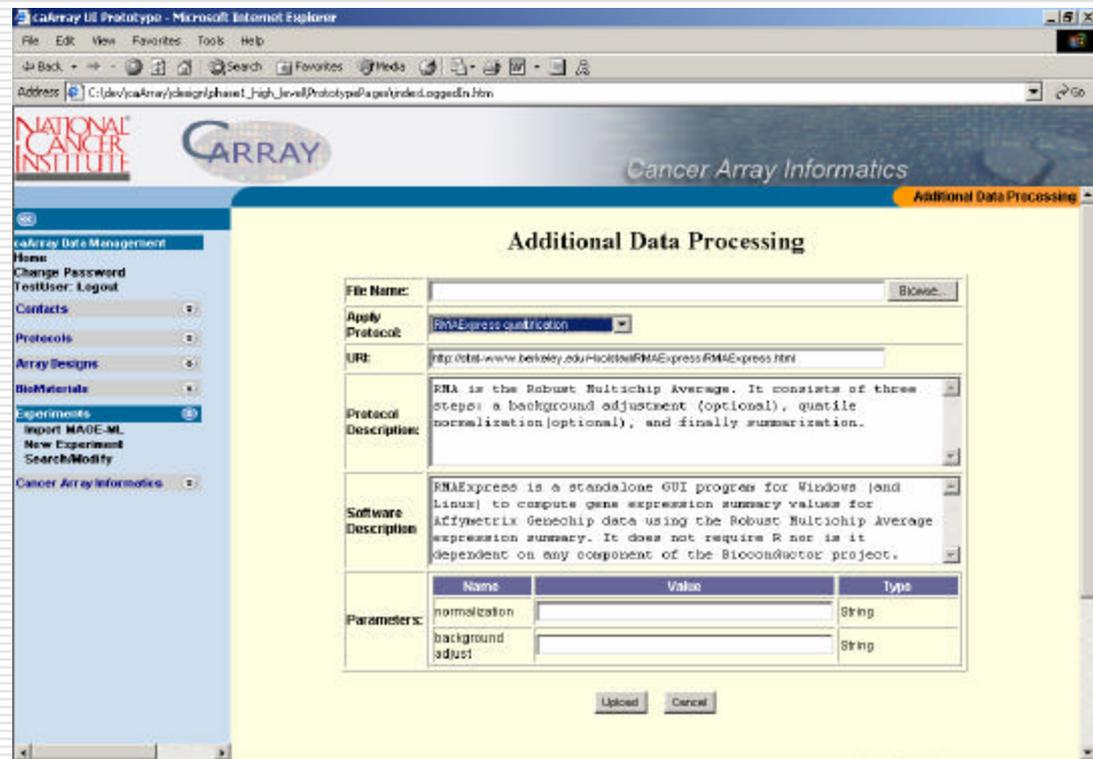
Image Acquisition Protocol:	Alymbix 418 scanning
Feature Extraction Protocol:	Alymbix MAS 5.0 Absolute Expression Analysis

Hybridizations (File Uploads)

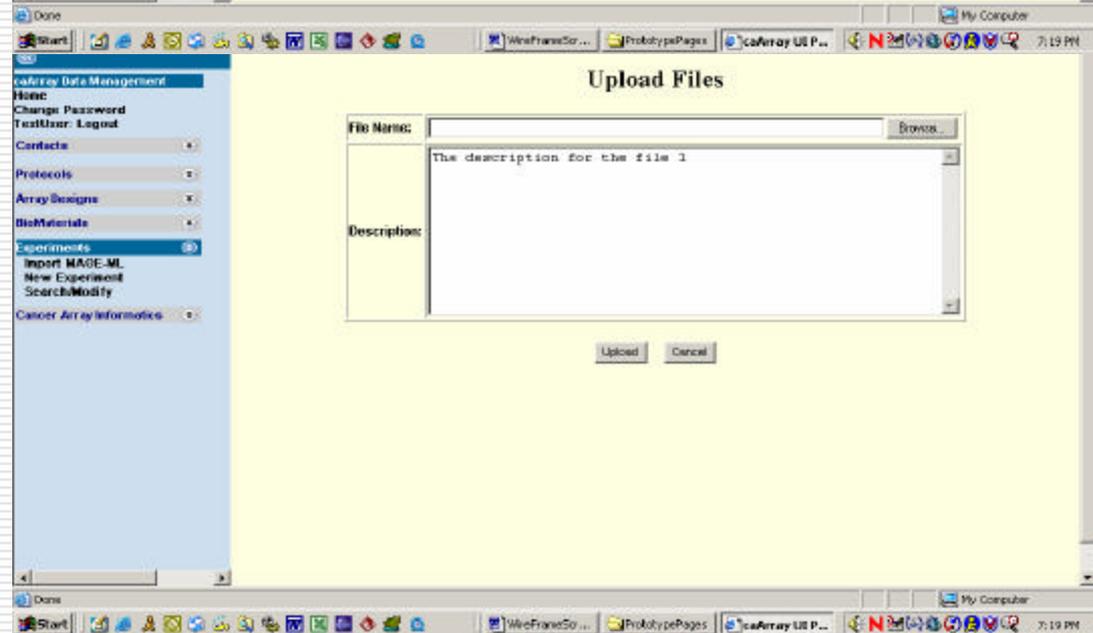
#	Array Batch	Array Identifier	Labeled Extracts	Factor Levels	Files to upload
1	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 1-12 Months Cell type: Cell type 1	cel <input type="text"/> Browse... exp <input type="text"/> Browse... bt <input type="text"/> Browse...
2	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 1-12 Months Cell type: Cell type 2	cel <input type="text"/> Browse... exp <input type="text"/> Browse... bt <input type="text"/> Browse...
3	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 13-24 Months Cell type: Cell type 1	cel <input type="text"/> Browse... exp <input type="text"/> Browse... bt <input type="text"/> Browse...
4	<input type="text"/>	<input type="text"/>	Extract: <input type="text"/> Spiked Control: <input type="text"/>	Age: 13-24 Months Cell type: Cell type 2	cel <input type="text"/> Browse... exp <input type="text"/> Browse... bt <input type="text"/> Browse...

[Upload](#)

Upload
processed data files



Upload
any other files



caBIG Compliance

General Principles

- Open access, open source
 - Both raw and processed data will be made available. Patient privacy and other concerns will be respected and addressed, but will not in and of themselves preclude sharing of appropriate data.
 - Software source code, use cases, designs, models, test plans, and other artifacts of the development process will be shared.
 - Where possible, open source enabling technologies will be selected
 - Good will among participants
 - caBIG Workspaces are collaborative teams, and participants are expected act accordingly.
 - Debate/discussion is encouraged, but professional demeanor and tone should be the norm.
- caArray allows submission and retrieval of all public data. The protection group/element concept in caArray allows researchers to preclude sharing of sensitive data.
 - All caArray use cases, design documents, test cases, source code and all other artifacts are being developed under an open source license. See caBIO license.
 - caArray utilizes only open source technologies such as MAGE-stk, JBoss, Xerces, Struts, Ant, OJB, etc
 - caArray is being built in an collaborative environment to address researchers' needs

caBIG Compliance

Architectural Principles

- Data and analytic services will be made available to caBIG using uniform application programming interfaces (APIs) and appropriate standard message formats.
 - The APIs and messages will be derived from common information models for biomedical data entities or objects, and will be accessible through interfaces that reflect these models.
 - APIs and messages will support the delivery of data and also of accompanying metadata, in order to ensure that aggregated data sets are comparable.
- caArray will allow programmatic interface to its data via the EJB Managers, MAGE-OM API, and the MAGE-ML document.
 - caArray is built upon MAGE-OM object model, MIAME and MGED-Ontology standards.
 - APIs and messages will support the delivery of data and also of accompanying metadata, in order to ensure that aggregated data sets are comparable.
 - caArray supports and extends MAGE-OM which allows for the deep annotation of microarray experiments according to MIAME.

caBIG Compliance

Architectural Principles, cont.

- Applications will be engineered to use the common caBIG APIs. → caArray is built to utilize and complement the caCORE infrastructure.
- Standards for data exchange formats will be used. → caArray is built upon the MIAME, MAGE-ML and MGED-Ontology standards.
- All systems, applications and selected standards will be documented. → caArray will include JavaDocs, a User Guide, and online Help.

caBIG Compliance Data Standards

- Data will be described by metadata elements that conform to an accepted standard such as ISO/IEC 11179. Metadata will be leveraged to achieve data interoperability and comparability.
 - caArray metadata will reside in caDSR, a ISO/IEC 11179 derived repository.
 - caArray is built to support MAGE OM (an OMG specification), as describe in Uniform Modeling Language (UML).

- Data and metadata will be constructed and drawn from appropriate vocabulary and ontology standards.
 - caArray utilizes the MGED Ontology, a set of open and standard controlled vocabularies and ontologies built to support annotation of microarray data.

- caBIG will consume of appropriate public, open access standards where they are available.
 - caArray is built upon MAGE-OM object model, MIAME 1.1, and MGED-Ontology standards. (See next slide)

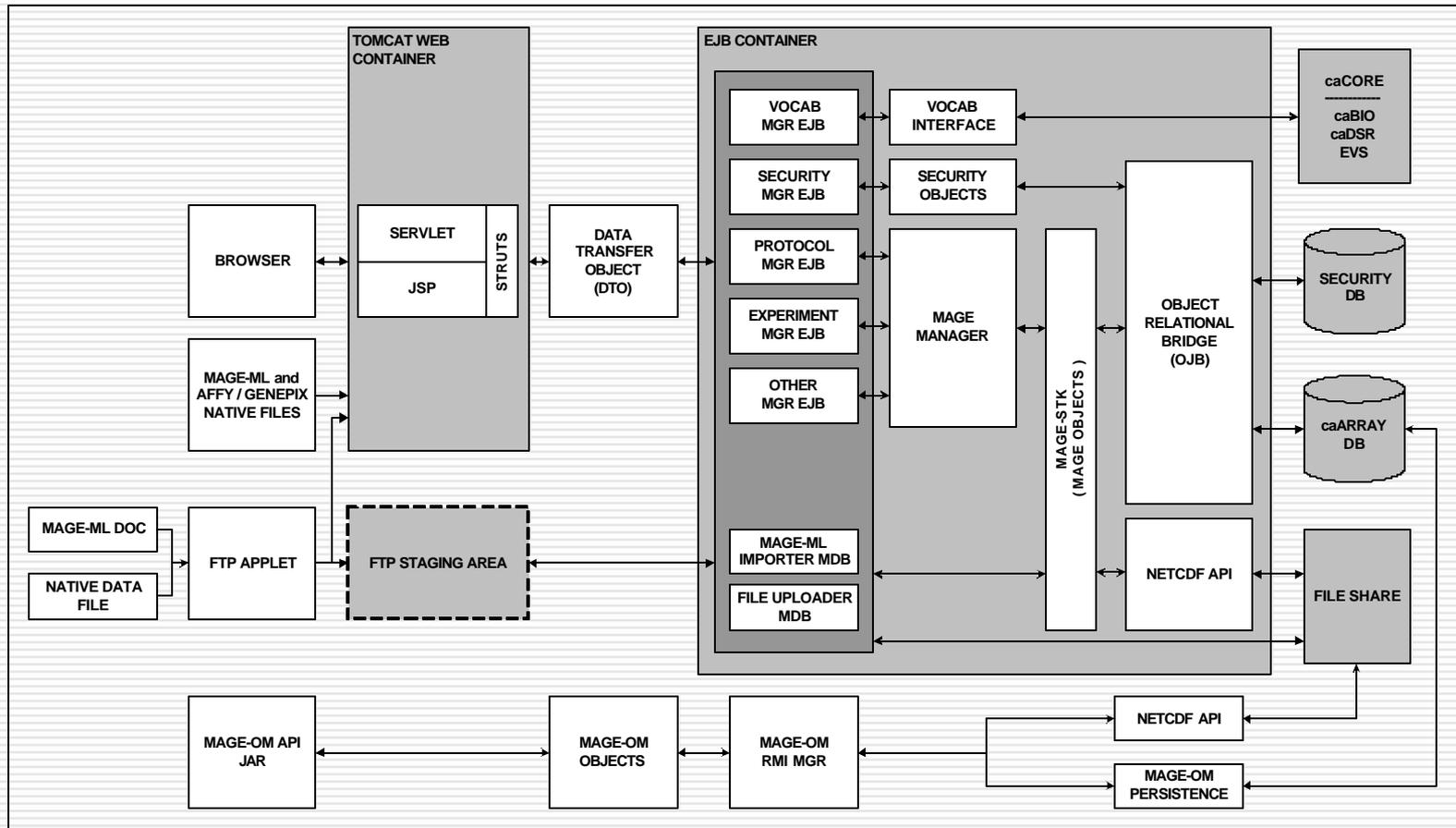
Compliance with International Standardization Efforts

- MIAME
 - Minimum Information About a Microarray Experiment
 - 1.1 Draft 6 (April 1, 2002)
 - http://www.mged.org/Workgroups/MIAME/miame_1.1.html
- MAGE-ML
 - MicroArray and GeneExpression Object Model and Markup Language
 - 1.1 (October 2003)
 - <http://www.omg.org/docs/formal/03-10-01.pdf>
- MGED Ontology
 - Microarray Gene Expression Data Ontology
 - 1.1.8 (April 2004)
 - <http://mged.sourceforge.net/ontologies/MGEDontology.php>

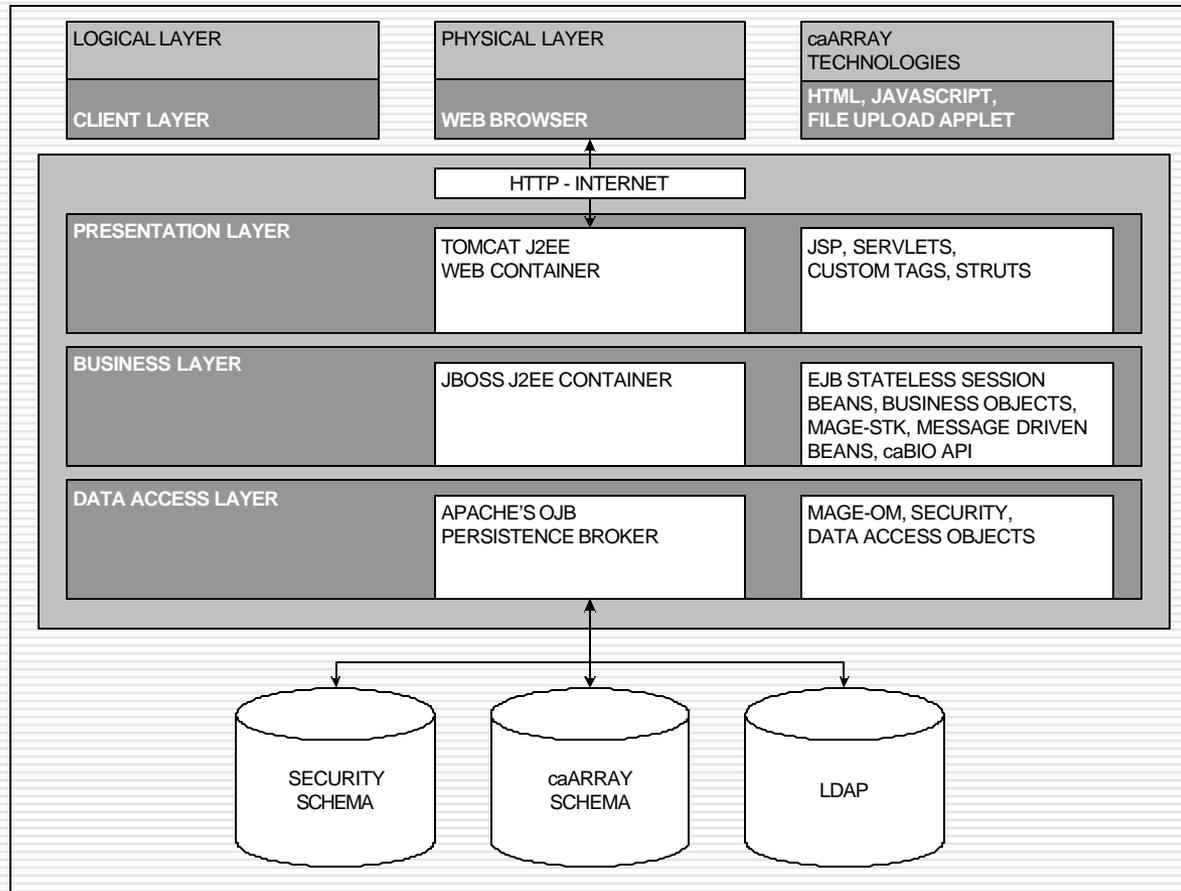
caArray Design/Architecture Considerations

- MAGE-ML Export/Import
 - Utilizes MAGE-stk 1.1 (evolving)
- User Access
 - Utilizes the Common NCICB Security Schema
 - Common NCICB authentication and authorization security module is in progress
- Scalable Robust Distributed Architecture
 - caArray is utilizing a both a J2EE web container and J2EE EJB container (JBoss implementation) for submission
 - Java Messaging Service (JMS) for asynchronous processing of large uploaded documents.
 - MAGE-OM API (RMI based) for retrieval
- Database Independence
 - caArray uses Apache's ObjectRelationalBridge (OBJ) as the Object Relational Mapping tool that allows transparent persistence for plain old Java Objects (POJO's) in the MAGE-stk toolkit against relational databases.
 - OBJ is also the current standard for caBIO allowing code reuse across projects.

caArray Architecture



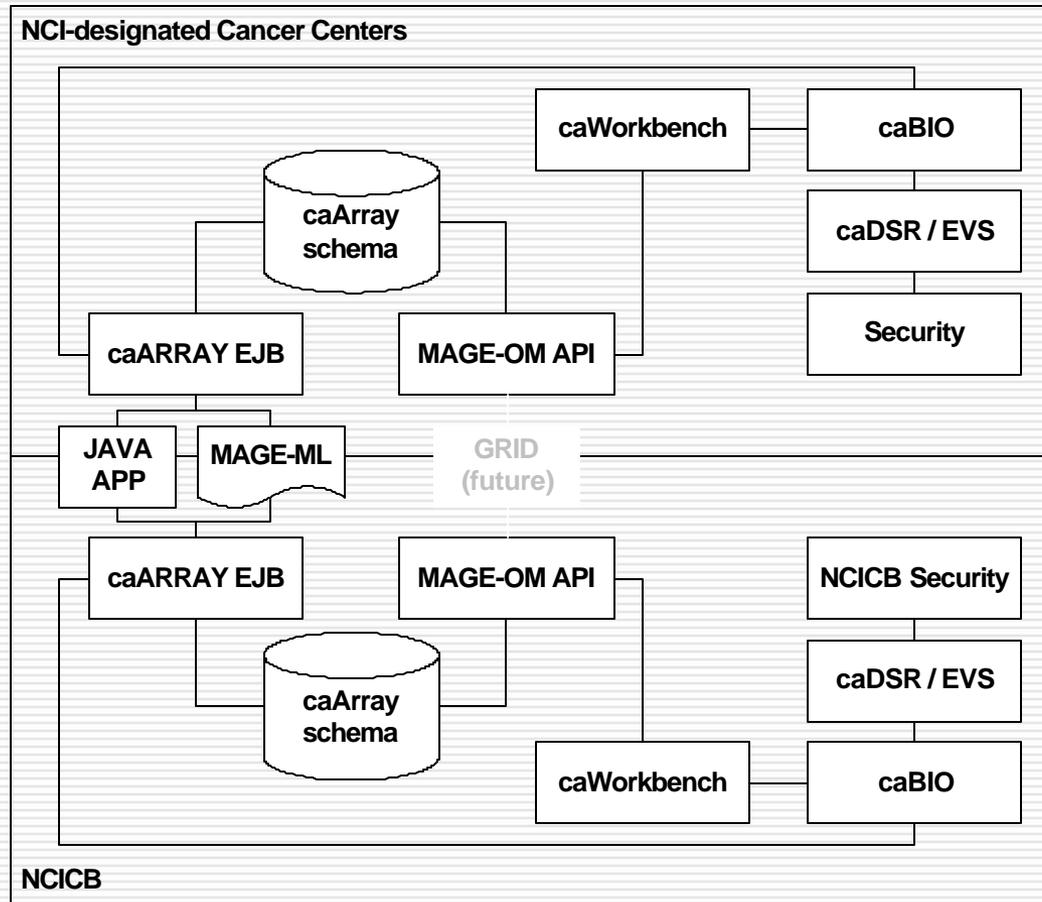
caArray Technologies



caArray Interfaces

- caArray provides access to its data via the following interfaces:
 - Programmatic
 - MAGE-OM API
 - caArrayEJB API via data transfer objects (DTO)
 - caArray Submission Portal
 - Document
 - MAGE-ML import/export
 - Native Affymetrix and GenePix file formats

Potential caArray Configuration



caArray Interfaces: Mage-OM API (cont'd)

- **MAGE-OM API** :Provides fine grain search and retrieval of all caArray data via a caBIO-like RMI based API.

For Phase I:

- The MAGE-OM API will be modified to map the MAGE objects to the new caArray database schema.
- RMI Security module will need to be incorporated and tested for user/group level data access.
- NetCDF API logic will need to be incorporated for faster retrieval of data

caArray Interfaces: Portal (cont'd)

- **caArray Submission Portal:** Via a user friendly GUI, caArray allows the users to submit, annotate and retrieve experiments.

For Phase I:

- MIAME 1.1 compliance
- Affymetrics & GenePix native file capabilities
- MGED Ontology for annotation
- MAGE-ML import/export

caArray Interfaces: caArrayEJB API (cont'd)

- **caArrayEJB API:** Provides transaction control, asynchronous processes, service location, common security and distributed capabilities for submission and retrieval of Microarray Experiments, MAGE-ML documents and its associated data files.

For Phase I:

- The caArrayEJB API will provide the above functionality via the caArray presentation layer.
- The caArrayEJB source code will be packaged and documented for installation at other cancer centers.
- The caArray team will only document how the caArrayAPI might be used for federated submission of micro array data.

caArray Interfaces: Federated (cont'd)

Future Phases:

- The actual proof of concept for federated submission of micro array data will be addressed during a future phase.

caArray Interfaces: Documents (cont'd)

- For MAGE-ML Export/Import:
 - caArray is planning to support both DataInternal and DataExternal .
 - The actual format of the DataExternal/ DataInternal is specified within the XML, such as tab delimited and others.
 - For Phase I, the Mage-ML import/export utility will only support the tab delimited format files such as output from Excel.

- Submission and retrieval of native Affymetrix and GenePix files

User Interface

- Conforms to NCICB UI Standards

The image displays two screenshots of the cancer.gov caARRAY user interface. The left screenshot shows a 'SPECIFY PERSON ATTRIBUTES' form with fields for First Name, Middle Initial, Last Name, Address, City, State, Postal Code, Toll Free Phone, Email Address, Fax Number, URL, and Affiliation. The right screenshot shows the 'CANCER ARRAY INFORMATICS' landing page with a magnifying glass over a microarray heatmap, a 'WELCOME TO THE caARRAY' section, and a 'LOGIN TO caARRAY' form with fields for LOGIN ID and PASSWORD.

Design Challenges and Solutions

- Presentation
 - GUI framework
 - Model View Controller 2
 - Flexible web page layout
 - Composite View

Design Challenges and Solutions

- Business
 - Data encapsulation
 - Transfer Objects
 - Common object to locate/lookup services
 - Service Locator
 - Abstract and decouple business services
 - Business Delegate
 - Encapsulate the DTO to MAGE-OM logic
 - Transfer Object Assembler
 - Uniform coarse-grained service access layer to clients
 - Session Facade
 - Process asynchronous MAGE-ML import & file uploads
 - Service Activator
 - Access to vocab/metadata services
 - Configurable Interface Pattern

Vocab/MetadataManager

Configurable Interface Pattern

- ❑ The VocabManager retrieves controlled vocabularies and metadata via an interface pattern.
- ❑ The implementations associated with this interface are configurable. This allows us to plug caArray into either the caCORE API, or a different metadata repository, or an XML metadata descriptor file.
- ❑ Via this interface, the VocabManager will allow the caArray application to perform attribute type checking, validation and population of enumerated lists and other controlled vocabularies.

Timelines for use of caDSR and creation of CDEs in caArray

- For phase I:
 - Configurable Interface Pattern used to abstract the source of controlled vocabularies and metadata behind the caArrayEJB API.
- On going:
 - Arrange to get the MAGE-OM API loaded into caDSR via UML loader
 - Looking into creation of CDEs for caArray GUI
- Future phase:
 - Addition of caArrayEJB APIs Data Transfer Objects (DTOs) into caDSR

Design Challenges and Solutions

- Persistence
 - Efficient representation of large data sets
 - NetCDF Data Capture Strategy
 - Decouple object and data source layer
 - Abstraction and Database Independence
 - Efficient materialization of objects
 - Lazy Materialization Pattern

NetCDF Data Capture Strategy

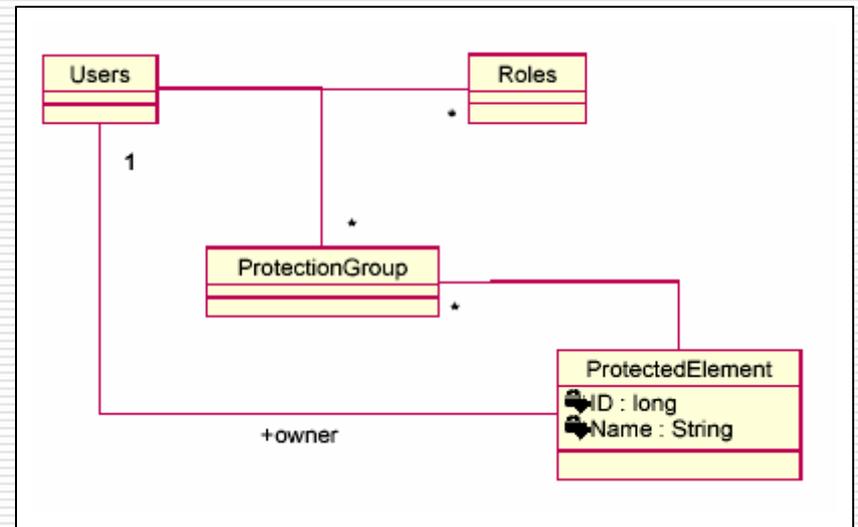
- ❑ Binary representation of numerical data.
- ❑ Easily manipulated by R and Java.
- ❑ Fast and efficient direct access to data matrices.
- ❑ Convertible to XML or tab-delimited forms
Java API can be written for transparent access.
- ❑ Self describing format.

Persistence Abstraction

- ❑ caArray uses Apache's ObjectRelationalBridge (OBJ) as the Object Relational Mapping tool that allows transparent persistence for plain old Java Objects (POJO's) in the MAGE-stk toolkit against relational databases.
- ❑ caArray utilizes OBJ's persistence facility to manage the connection to the data source, store, and restore MAGE-stk objects.
 - The mapping of MAGE-stk objects and their relationships to the database entities is described in OBJ repository files to allow OBJ to know how to persist or restore them.
- ❑ OBJ is also the current standard for caBIO allowing code reuse across projects.
- ❑ Benefit:
 - Decoupling between Object and Data source layers

User Access Roles

- **User Manager**
 - A user who is responsible for creating data sharing consortiums, users, organizations and people.
- **Configuration Manager**
 - A user who is responsible for managing the protocols, hardware, software and array designs.
- **Ontology Manager**
 - A user who is responsible for managing dynamic vocabularies.
- **Experiment Manager**
 - A user who is responsible for annotating biomaterials and maintaining the data associated with experiments.
- **Curator**
 - A user who is responsible for the integrity of the data for a particular group or a set of groups.
- **Repository Curator**
 - A user who is responsible for the integrity of the data stored in the repository.
- **Data Owner**
 - Any user that is currently listed as the owner of any protected data element in the system.
- **User**
 - Any individual who will use the system to submit data or search through and utilize existing data.



Data Models

□ caArray data model

- The data model was derived by annotating the MAGE-OM API with Xdoclet tags.
- The Java files were processed by an Xdoclet module (Apache's Torque).
- The process generated the OJB repository.xml file and the SQL DDL schema.
- The schema was then optimized for performance.
- Benefit: Using a code generator, the schema can be regenerated when the MAGE-OM changes.

□ Security data model

- Utilizing NCICB common security data model.

Proof of Concept

- Developed Protocol Management reference implementation (proof of concept) that implemented end-to-end functionality for protocol management screens.
- Using the proof of concept, one can perform the following:
 - Search, Add, Update, and Delete Protocols
 - Add Hardware/Software/Parameter to Protocol

Wrap Up

Questions?

<http://caarray.nci.nih.gov>