**NCI caBIG®**

**Common Biorepository Model**

**CBM Information**

**CBM 1.0Beta Test Grid Installation**

**December 2010**

**NCI CBIIT – CBM Team**

[1.0 Common Biorepository Model Background, FAQ 3](#_Toc275722390)

[1.2 What does CBM 1.0Beta entail? 3](#_Toc275722391)

[1.3 Why should CBM1.0Beta (October 2010) be tested by an institute/vendor? 4](#_Toc275722392)

[1.4 How will my institute’s CBM grid service be used? 4](#_Toc275722393)

[1.5 How will having a CBM grid service at my institution help me share data and specimens? 5](#_Toc275722394)

[1.6 How will CBM be supported – responsibilities (DRAFT) 5](#_Toc275722395)

[1.7 How were the vocabulary lists generated? 6](#_Toc275722396)

[2.0 Overview CBM 1.0Beta Testing Grid Service Package 6](#_Toc275722397)

[2.1 What does CBM 1.0Beta Test Grid Service Package contain? 6](#_Toc275722398)

[3.0 ETL into CBM database 7](#_Toc275722399)

[4.0 Instructions for Deploying CBM service at your test site/institute. 7](#_Toc275722400)

[4.1 Pre-requisites 7](#_Toc275722401)

[4.2 Unzip the CBM deployment package 7](#_Toc275722402)

[4.3 Set up MySQL database 7](#_Toc275722403)

[4.4 Set up your JBoss Container 8](#_Toc275722404)

[4.5 Deploy Grid service 8](#_Toc275722405)

[4.6 Start JBoss Server 10](#_Toc275722406)

[4.7 Run a test query (script) 10](#_Toc275722407)

[5.0 Testing the CBM Grid Services that are up and running (next steps) 12](#_Toc275722408)

[5.1 Looking for CBM Service over the Training caGrid Portal – 12](#_Toc275722409)

[6.0 CBM Help Resources 20](#_Toc275722410)

# 1.0 Common Biorepository Model Background, FAQ

For goals of CBM and the latest on the CBM Challenge, please refer to the CBM website: <https://cabig.nci.nih.gov/workspaces/TBPT/CBM/>

SVN storage location:

[https://ncisvn.nci.nih.gov/svn/common\_biorepository\_model/trunk/](https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/caCORE_SDK/models/)

CBM1.0Beta is a another test package containing revised model, service files for standing on the training caGrid, and fixed vocabulary terms that institutions may start mapping their data to the CBM1.0Beta database, a step needed in creating a CBM service at an institute, and published on the grid.

CBM1.0 will be developed by the “Specimen Resource Locator”Developers (to be awarded in 2010 Q4). CBM1.0 will take the CBM1.0Beta model and build up the ECCF/SAIF documentation and adjustments required for caBIG® conformance. They will also develop the test suite that will be used for caBIG® conformance/compliance testing.

## 1.2 What does CBM 1.0Beta entail?

This is the latest iteration of the CBM model to provide sites for testing by the biospecimen management systems (vendor solutions, caTissue Suite, and in-house institute-developed solutions) and provide feedback on the model, service, and test with NCI. In this version:

* The vocabularies that describe the specimen types, diagnoses types, etc. have been incorporated into the NCI Thesaurus (required for caBIG® semantic interoperability) and are fixed for at least 1 year (until May 2011 or later).
* The CBM1.0Beta Model vocabulary has been reviewed through caDSR, and final adjustments have been made to ensure caDSR can record the model – this will be helpful as NCI moves to the official CBM1.0 release (expected 2011).
* caBIG compatibility for this version is still under evaluation, as NCI is re-defining compatibility terms under Services-Aware Interoperability Framework (SAIF) and sub-framework—the Enterprise Conformance and Compliance Framework (ECCF).
* This is the version NCI would like institutes (vendors and biorepostories) to start testing and provide feedback on.
* Establish grid nodes from COTS and custom systems
* CBM 1.0Beta EA model is found in the CBM package and can also be found at: [https://ncisvn.nci.nih.gov/svn/common\_biorepository\_model/trunk/caCORE\_SDK/models/CBM with Value Domains.EAP](https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/caCORE_SDK/models/CBM%20with%20Value%20Domains.EAP)

## 1.3 Why should CBM1.0Beta (October 2010) be tested by an institute/vendor?

* The CBM 1.0Beta version has the vocabulary lists fixed for the near future to be part of CBM1.0 – the list will be fixed at a minimum through October 2011.
* An early limited release CBM1.0Beta-May2010 was released to those vendors who were willing to test with NCI and had test data (not based in real repository information). These vendors provided additional input into the model and instructions, and NCI has revised the model, and incorporated values into the offered CBM.sql database for ease of the Extract-Transform-Load process.
* CBM 1.0Beta will be provided as a baseline for the CBM1.0/SRL developers to create a caBIG® compliant service (under the Specimen Management set of services),develop a “TEST SUITE” for validation, and revise the Specimen Resource Locator to connect to the CBM services to allow for up-to-date biorepository collection information to be provided to researchers (expected 2011).
* We advise you to use this version and this vocabulary list to start ETL process against your vendor or home-developed management system.
* **NOTE:** CBM 1.0Beta is not yet fully designated to be a caBIG® certified/compliant model, as this is still in the process of being defined in NCI-CBIIT. We will let the community know when this happens, and the impact it may have in your ETL process.

## 1.4 How will my institute’s CBM grid service be used?

* Rudimentary (not user-friendly) – it will immediately be queryable over the caGrid (for these Testing iterations – we suggest all CBM services to be posted on the caGrid Training Portal – instructions below).
* The ultimate NCI consumer of these services will be the revised NCI Specimen Resource Locator (SRL) -- (SRL: <http://biospecimens.cancer.gov/locator>), a search tool made available to the research community for locating biospecimens. The NCI SRL pools information on the specimen collections of participating biobanks and biorepositories into a single, searchable database, allowing researchers to make database queries for biospecimens, and based on the results directs them to the biobanks that match their requests. The new Specimen Resource Locator work will begin in early 2011, and will use CBM1.0 grid services.
* Vendors/institutes may also develop their own web-interface to query their and other CBM grid services.
* In the short term while the SRL is undergoing changes, NCI will work to identify other methods to query the Grid portal in a more user-friendly manner (for demonstration purposes and use as a development tool for testing queryability of the model for SRL use). This include caBIG® caB2B or other technologies for displaying grid-queries. We will share these examples with the community so the vendors/institutes may also use them when demonstrating your efforts with CBM.

## 1.5 How will having a CBM grid service at my institution help me share data and specimens?

* See section 1.4 above, and, in addition:
* Regardless of the data systems used at your institution you can use CBM to share summary level de-identified information about each biobank.
* You will be listed in the revised Specimen Resource Locator; a guide for researchers to find specimens.
* The Specimen Resource Locator will direct people to your institute (via the contact person name you provide). At that point, the researcher seeking specimens will discuss with the biorepository representative, to find out more about the specimens and what data sharing agreements, material-transfer agreements are in place.

## 1.6 How will CBM be supported – responsibilities (DRAFT)

* Biospecimen Management System Vendor/In-house biospecimen management system developer:
  + Will know how to stand up a grid node for CBM on the test caGrid
  + Recognize how to do the ETL mapping of the terms in their proprietary system to the CBM model terms
* Bioinformatics/IT Department at Institute:
  + Will provide a space on their server which will serve to host the CBM database (accessible to outside world – in their DMZ)
  + Work with Biospecimen Management System Vendors
* NCI caBIG® will have a set of resources to provide help as institutes stand up a CBM service that may include:
  + CBM-team [**NCIcbm@mail.nih.gov**](mailto:NCIcbm@mail.nih.gov) **(currently active)**
    - Notify if you are using the CBM for NCI to note and publish who is working on the CBM initiative (even in early testing).
    - Point-of-contact for general CBM questions – this team will redirect questions to appropriate NCI-support.
    - Specimen Resource Locator questions and participation requests.
  + TBPT-KC – Forums for CBM end-users and institutes standing up a grid CBM service (currently active)
  + caGrid-KC – help with caGrid-specific questions (currently active)
  + CBM-SRL Developer Team (not active, we will let community know what level of help and how to contact the support when process is defined).

## 1.7 How were the vocabulary lists generated?

* + The vocabulary for the CBM was developed with the help and input from the Specimen Resource Locator stakeholders, NCI and other NIH institutes leaders in the biospecimen, Biorepository and pathology fields. Having finalized a list of values for Anatomic Sources, Diagnoses, Preservation Methods and Specimen Types, the list was passed on to the Enterprise Vocabulary Services team at the NCI whose lexicologists provided them with NCI approved definitions, synonyms, and NCI Thesaurus codes. Once the terms were added to the NCI Thesaurus, the terms were subsequently added to the NCI metaThesaurus where they were mapped to widely used scientific standards, notably ICD9, ICD10 and SNOMED thus allowing them to be cross-referenced by CBM implementers.
  + The NCI/NIH stakeholders included: Office of Biorepositories and Biospecimen Research; National Heart, Lung, and Blood Institute; Division of Cancer Control and Population Sciences; Cancer Diagnosis Program (CDP); National Institute of Diabetes and Digestive and Kidney Diseases; Office of Rare Diseases Research.

# 2.0 Overview CBM 1.0Beta Testing Grid Service Package

## 2.1 What does CBM 1.0Beta Test Grid Service Package contain?

* The service files needed for CBM 1.0Beta to be tested at your institute/vendor site.
* The CBM 1.0Beta EA model (*CBM with Value Domains.EAP*, *CBM with Value Domains.xmi*)
  + Vocabulary lists –all now included in the NCI MetaThesaurus (required for semantic/syntactic compatibility) and with additional correlations against known codes (SNOMED, ICD-9, etc).
    - Values are also incorporated in the EA model attributes.
  + The model can be viewed, traversed through via HTML representation of the EA model, found here:
  + HTML view of EA model: https://ncisvn.nci.nih.gov/svn/common\_biorepository\_model/trunk/html\_documentation/index.htm
* *EVS definitions and mappings\_caDSR.xls* - Excel file that contains mappings to SNOMED, ICD, and synonyms found in NCI Thesaurus for the terms used in CBM1.0Beta
* *CBM.SQL*
  + MYSQL data file that can be used for ETL institute data into. This schema has the expected values for each class in the model.
  + NOTE: What is not defined in the CBM.SQL populated file is the SpecimenCollectionSummary.patientAgeGroupAtCollection: this should be defined in 5 year increments, from 5 to 85, 200
    - 5, 10, 15, 20…85, [86-200]
* Document CBM next steps and answers to frequently asked questions (this document).
  + Instructions for how to Stand up a CBM Service Grid node on the Training Grid
  + Resource List – for contact names and relevant websites to use for references

# 3.0 ETL into CBM database

* We have included a “populated” CBM database at:

<https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/database/CBM.sql>

* This database is CBM schema, “populated” with the diagnosis list of values and the other terms required for reference.
* This serves as a CBM database schema that CBM-participants can map their data into via ETL (Kettle or other preferred ETL programs).

# 4.0 Instructions for Deploying CBM service at your test site/institute.

## 4.1 Pre-requisites

1. Install JDK 1.5.0\_14 or later (NOT JDK 6) (i.e. JDK 1.5.0u22 package that installs JDK 1.5.0\_22)
   1. Set JAVA\_HOME environment variable to point to your JDK directory.
2. Install Ant 1.7.0 (Note: Make sure it is this version)
   1. Set ANT\_HOME environment variable to point to your Ant directory.

## 4.2 Unzip the CBM deployment package

Set GLOBUS\_LOCATION environment variable to point to the CBM\_DIR/ws-core-4.0.3 directory

CBM\_DIR is the directory where you unzipped the package.

## 4.3 Set up MySQL database

1. If you are using a platform other than Windows or Mac OS X, please refer to the following instructions to read all table names as lower case: <http://dev.mysql.com/doc/refman/5.0/en/identifier-case-sensitivity.html>
2. Verify you are using MySQL 5
3. Create a MySQL database (use “cbm” if possible)
4. Create a user that has permissions to access the database: <http://dev.mysql.com/doc/refman/5.1/en/adding-users.html>
5. Load the test data into the mysql database
   1. Example using mysql command prompt: source cbm\_data1.sql

Note: A simple database with limited data will be included and will be included in the CBM site. One can use CBM.SQL to create a cbmX.database and populate sample data using MYSQL programs.

## 4.4 Set up your JBoss Container

1. Copy the container/jboss-4.0.5.GA directory to a directory of your choice.
2. Set your JBOSS\_HOME environment variable to point to the jboss-4.0.5.GA directory
3. Edit the JBOSS\_HOME/server/default/deploy/jbossweb-tomcat55.sar/server.xml
   1. Change the Connector port to the value for your server
   2. Default is <Connector port="8080"

Edit JBOSS\_HOME/server/default/deploy/wsrf.war/WEB-INF/web.xml

* 1. Change the default port parameter to the value for your server
  2. <init-param>
  3. <param-name>defaultPort</param-name>
  4. <param-value>8080</param-value>
  5. </init-param>

## 4.5 Deploy Grid service

1. Copy the “cbm” folder (Grid service) to a folder of your choice.
2. Copy the cbm/lib/**CBM-orm.jar** to a folder of your choice
3. Update the jar
   1. Extract the jar (for extraction and creating jar files, the \Java\jdk1.5.0.5.0\_22\bin\jar program may be used)
   2. Update the following properties in the file **hibernate.cfg.xml** to point to the database with data (in the test package, “cbm3”)

<property name="connection.driver\_class">com.mysql.jdbc.Driver</property>

<property name="connection.url">jdbc:mysql://localhost/cbm3</property>

<property name="connection.username">user</property>

<property name="connection.password">password</property>

* 1. Re-create/update the jar with the new file

**Notes:** This current CBM service does not authenticate users.  Users will not  be prompted for  username/password.  The username/password referenced above is for the CBM  database.   That information will be used by the service to connect to the   database when queries are submitted to it.  Localhost is used with the assumption that the CBM grid service is running on the same machine as the CBM  database.  If the service and database are running on separate servers, the hostname of the database server will need to be used.

(i.e. in the cbm-orm directory, run the command: …/jdk1.5.0\_22\bin\jar –cvf cbm-orm.jar \*)

1. Copy the updated jar to the cbm/lib directory
2. For Grid-Node Testing, Setup the “Institution Name, etc” as it will be known on the Grid Service.
   1. Edit cbm/etc/ serviceMetadata.xml
   2. Change the PointOfContactCollection, should also be changed/populated.

<ns2:pointOfContactCollection>

<ns3:PointOfContact affiliation="CBM TestSite”email="cbm\_testsite\_poc@yyy.com" firstName="TesterFirstName" lastName="TesterLastName" phoneNumber="" role="Developer" xmlns:ns3="gme://caGrid.caBIG/1.0/gov.nih.nci.cagrid.metadata.common"/>

</ns2:pointOfContactCollection>

* 1. Change the associated hostingResearchCenter to correspond with data associated with your institute:

<ns1:hostingResearchCenter>

<ns10:ResearchCenter displayName="CBM Test University" shortName="TestUniv" xmlns:ns10="gme://caGrid.caBIG/1.0/gov.nih.nci.cagrid.metadata.common">

<ns10:Address country="US" locality="Columbus" postalCode="44333" stateProvince="OH" street1="460 W 12th Ave" street2=""/>

<ns10:ResearchCenterDescription description="" homepageURL="" imageURL="" rssNewsURL=""/>

This “shortName” will be displayed on the Training Grid Portal (see section 4.1).

1. Update the index service registration URL
   1. Edit cbm/deploy.properties file
   2. Change the following line to the URL of your index service, depending on whether you are indexing to the caGrid Training Grid or the caGrid Production Grid. (NOTE: do not change the “\” characters): index.service.url=http\://index.training.cagrid.org\:8080/wsrf/services/DefaultIndexService
   3. Note: Since we are still testing CBM1.0Beta, it is okay to keep the association to the Training Grid

**Notes:** The    index.service.url refers to the central naming service for the  grid.   For the training grid, everyone uses  index.training.cagrid.org.   The  production grid uses a  different url.  When a service  starts up, it  connects to  the index service to register itself.   Others can query the  index service to get a list of available  services.  The  portal ( <http://portal.training.cagrid.org>)    is a web-app that interfaces with the index service to display  active   services.

* 1. Note: the caBIG Production Grid Index service URL is: <http://cagrid-index.nci.nih.gov:8080/wsrf/services/DefaultIndexService>

1. Deploy the service to your JBoss container (from the CBM\_DIR directory (…/cbm)): ant deployJBoss
2. (NOTE: ant undeployJBoss; ant deployJBoss –logfile XXX.txt to record output)
3. Note: you have to be in the ../cbm/directory (where build.xml is hosted)

When a vendor/CBM participant  deploys  the service at their site,  the service will register  with index  service.  Once the service has  registered,  it’ll appear on the  Training Grid Portal (<http://portal.training.cagrid.org>).     Other users will be able to submit queries via the  portal or use  the  published service url’s to connect their own  clients.  A  service does not  need to register with the  index to be used.  If  the caller already knows  the  hostname or ip address, they can call the  service  directly.

## 4.6 Start JBoss Server

1. $JBOSS\_HOME/bin/run.sh (or run.bat)

(In Windows: %JBOSS\_HOME%/bin/run)

1. Open the following address in a web browser: <http://localhost:8080/wsrf/services/cagrid/CBM>
2. You should see a message similar to “Hi there, this is an AXIS service!”

## 4.7 Run a test query (script)

1. Change to the CBM directory
2. Run the runQueryPerson.sh script
   1. Note: if you are using Windows, run the ant inside the file command directly
   2. To run other queries, simply change the xml file that is reference in the command
3. How many Participants have specimens from anatomic source: “Prostate Gland”

XML query (saved as prostateQuery.xml:

<ns1:CQLQuery xmlns:ns1="http://CQL.caBIG/1/gov.nih.nci.cagrid.CQLQuery">

<ns1:Target name="gov.nih.nci.cbm.domain.LogicalModel.ParticipantCollectionSummary">

<ns1:Association name="gov.nih.nci.cbm.domain.LogicalModel.SpecimenCollectionSummary" roleName="provides">

<ns1:Attribute name="anatomicSource" predicate="EQUAL\_TO" value="Prostate Gland"/>

</ns1:Association>

</ns1:Target>

</ns1:CQLQuery>

The query is run by running on the cmd line the shell script “runQueryParticipantProstate.sh (or copying the content file at the command line >ant –f ….)

# 5.0 Testing the CBM Grid Services that are up and running (next steps)

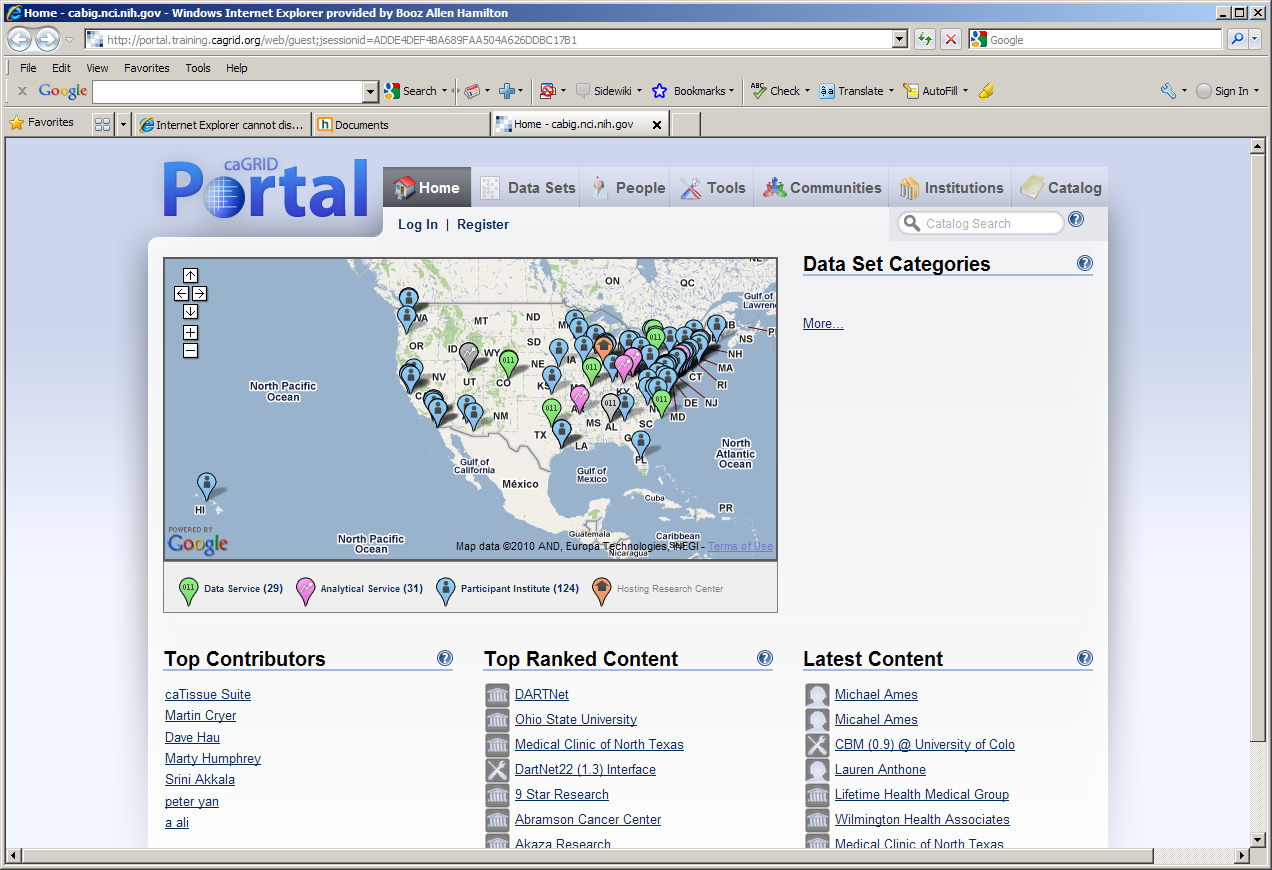
* Please notify NCI ([NCIcbm@mail.nih.gov](mailto:NCIcbm@mail.nih.gov)) when your vendor/institute site. This will notify NCI to see that your service is up and running for this test phase.
* **NOTE: By the end of 2011, there should be a set of “TEST CBM GRID SUITE SCRIPTS” that will enable anyone to test their service against prescribed scripted queries, to ensure that the CBM service your institute stands up is up and queryable.**
* **For now, we will use the caGrid Training Portal to do some simple testing and querying.**

## 5.1 Looking for CBM Service over the Training caGrid Portal –

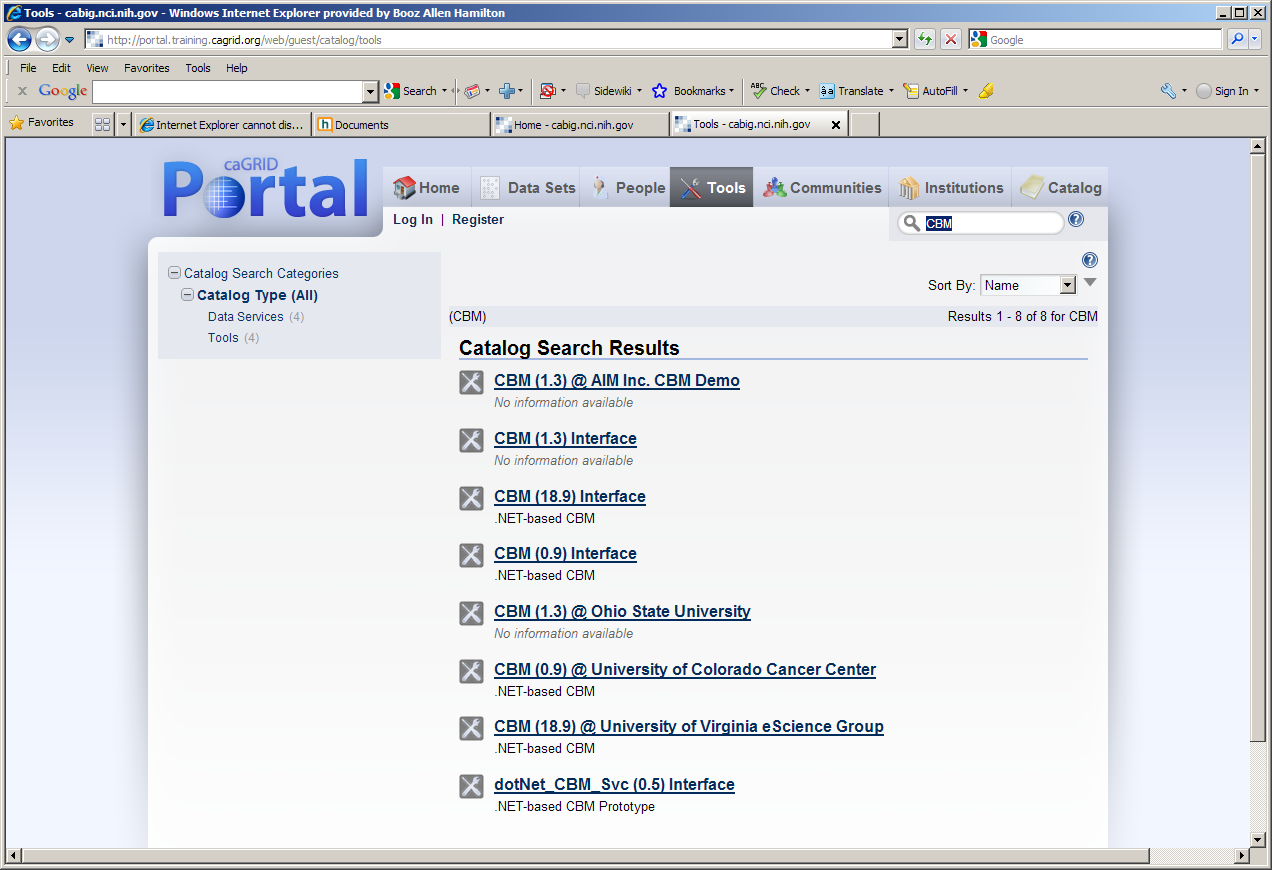
* You will be able to query the grid on the Training Grid portal:

How to find the grid service:

<http://portal.training.cagrid.org/web/guest/home>

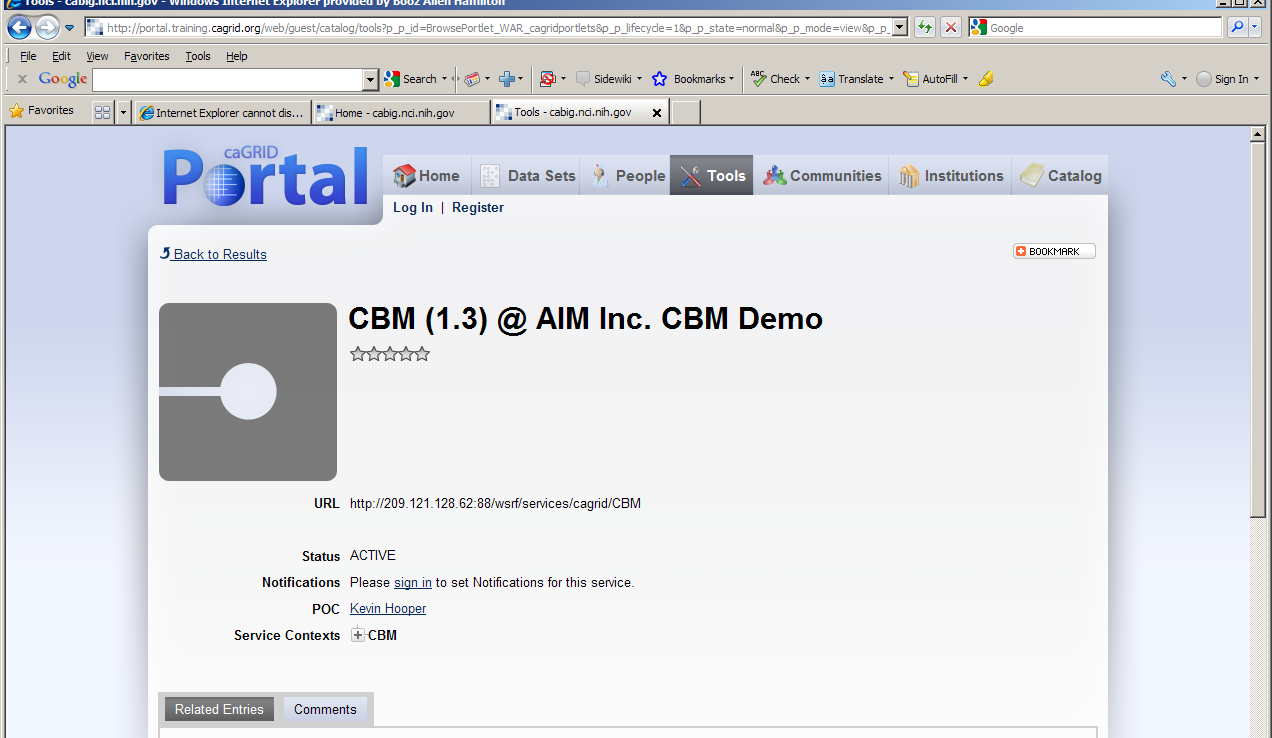


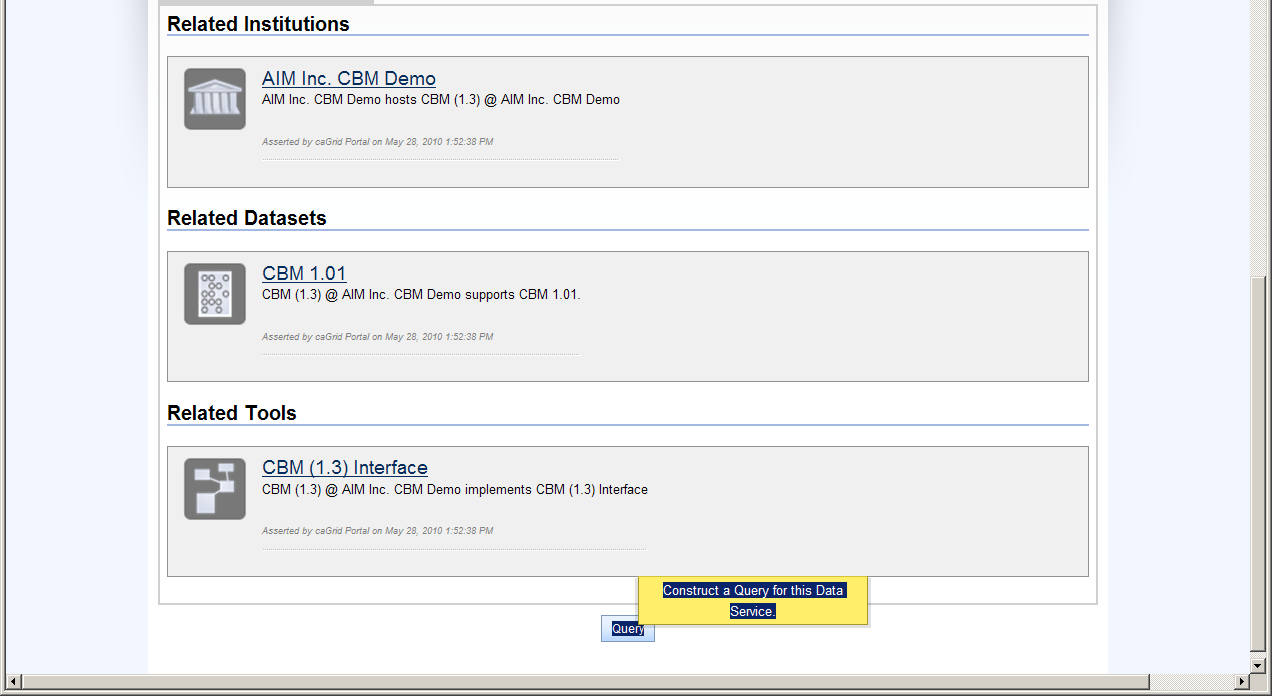
Go to top bar and click on “Tools” and then in the “Catalog Search”, enter “CBM”



Your service should be up within 15 minutes, and shown here.

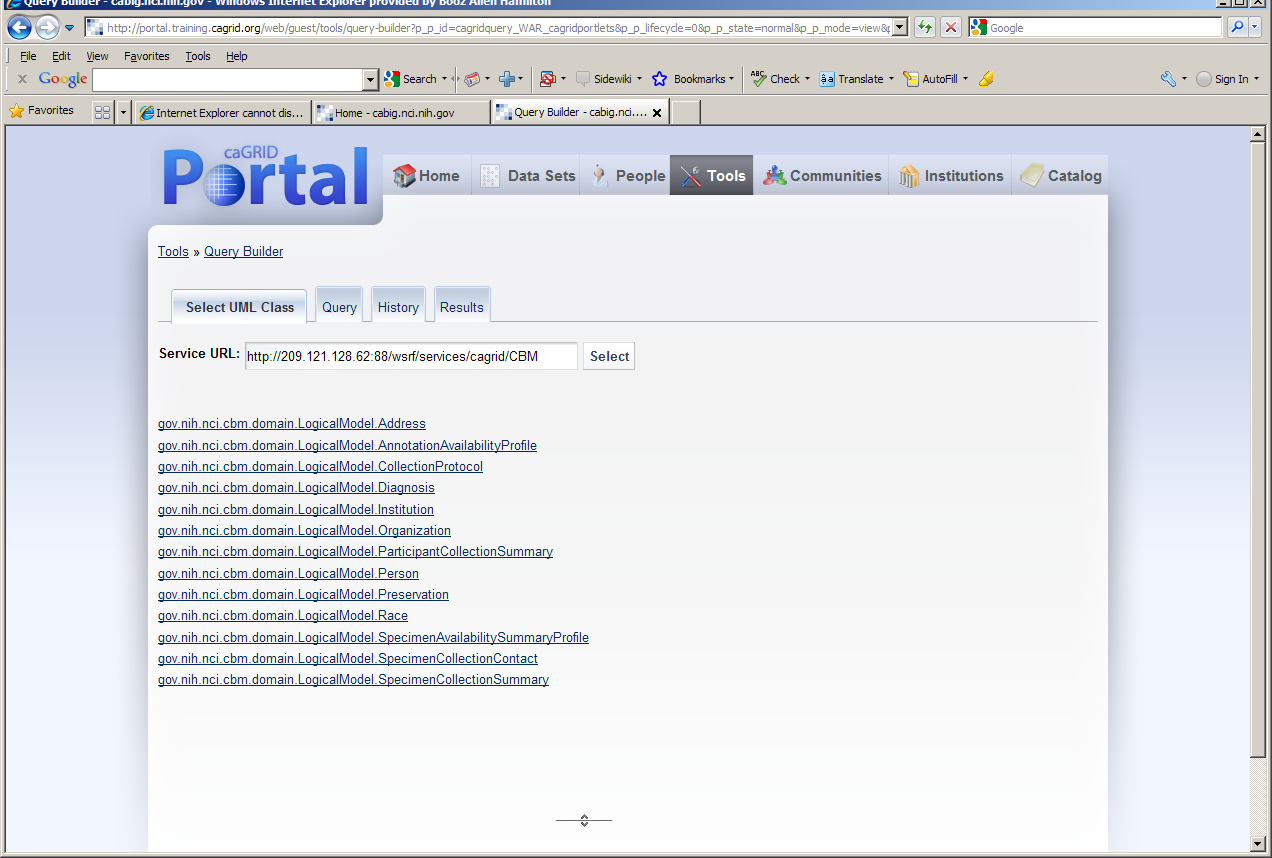
Clicking on the service link:



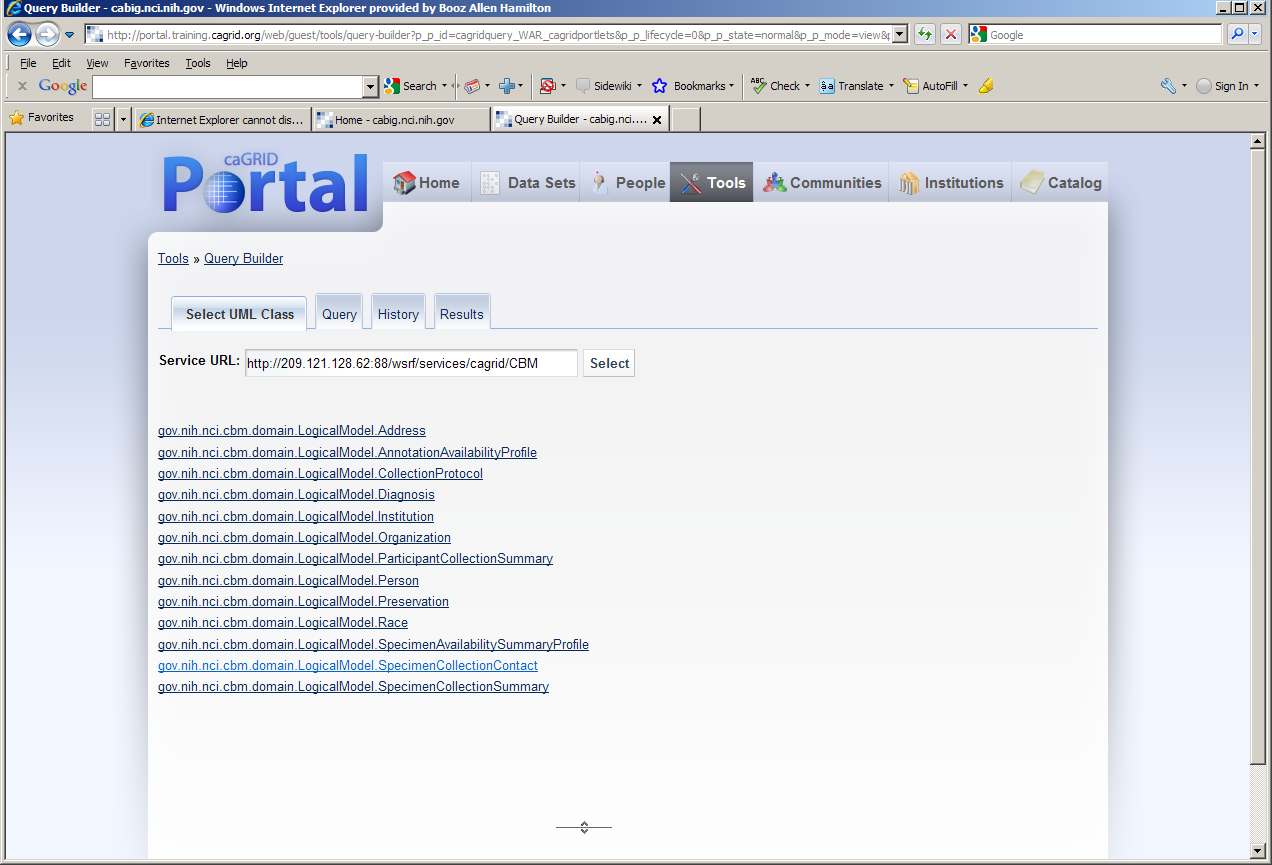


Simple Query:

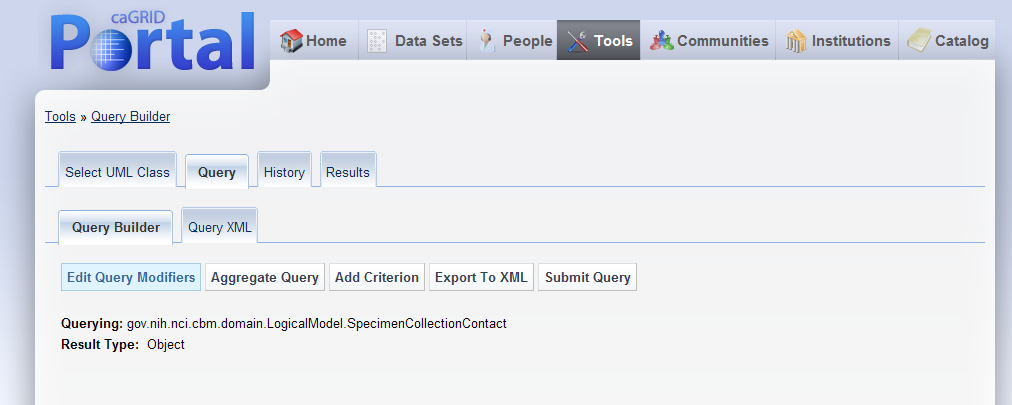
* Once you select your service and press “Query”) , you can select a UML class : i.e. “ …. CollectionProtocol” and query the count (default query modifier setting):

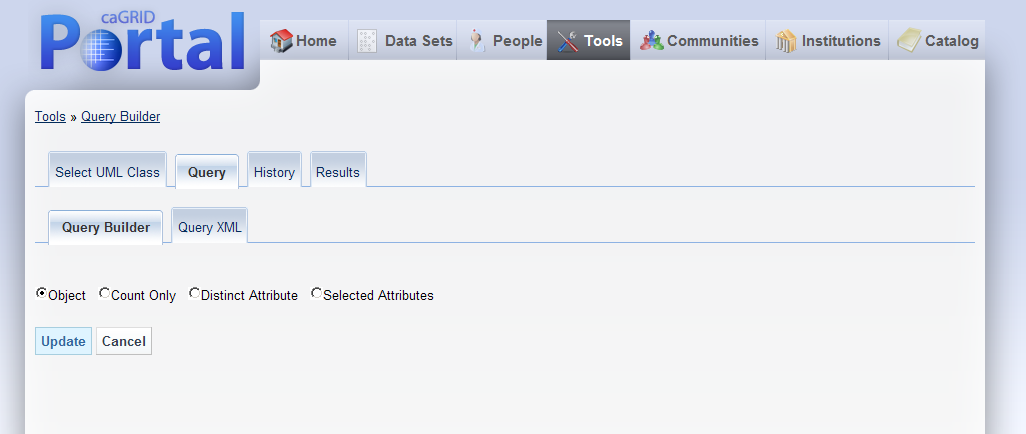


One can traverse through the model to select and build a query. A simple query is on the SpeicmenCollectionContact:

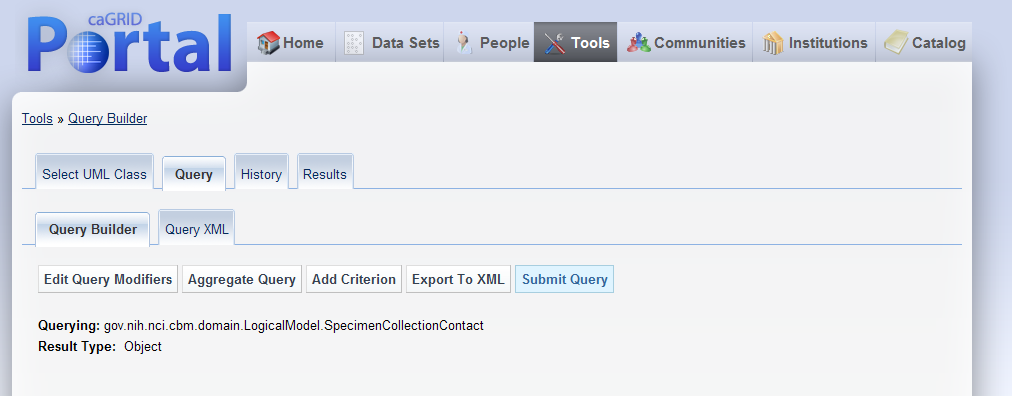


Click on that link, and then also make sure the entire Object is returned, so you must “modify query”:

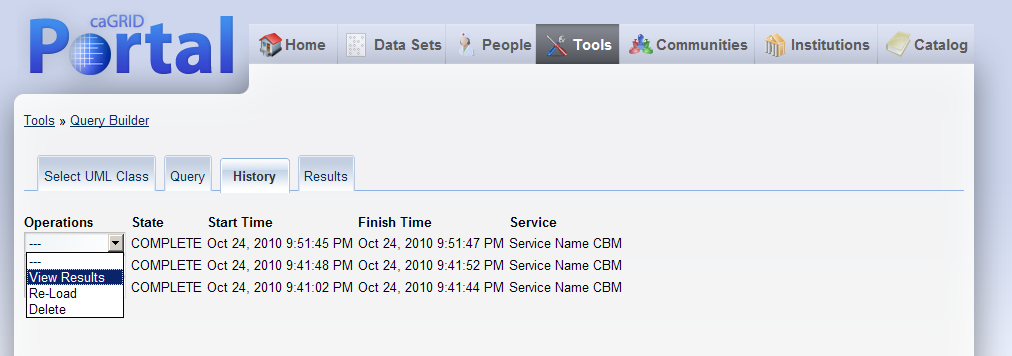


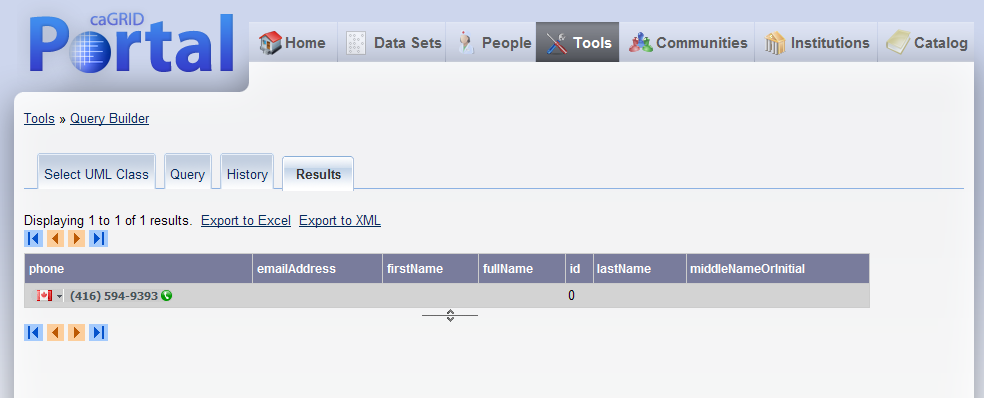


Hit Update, and then “Submit Query Tab”

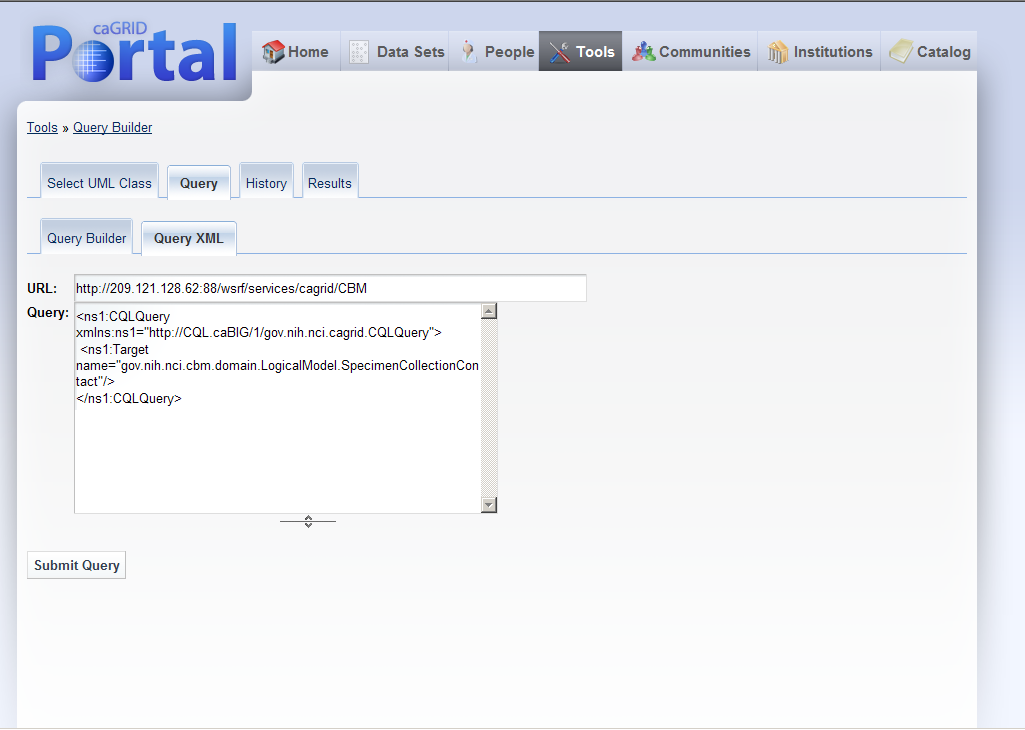


A screen should appear, and you can select “View Results”

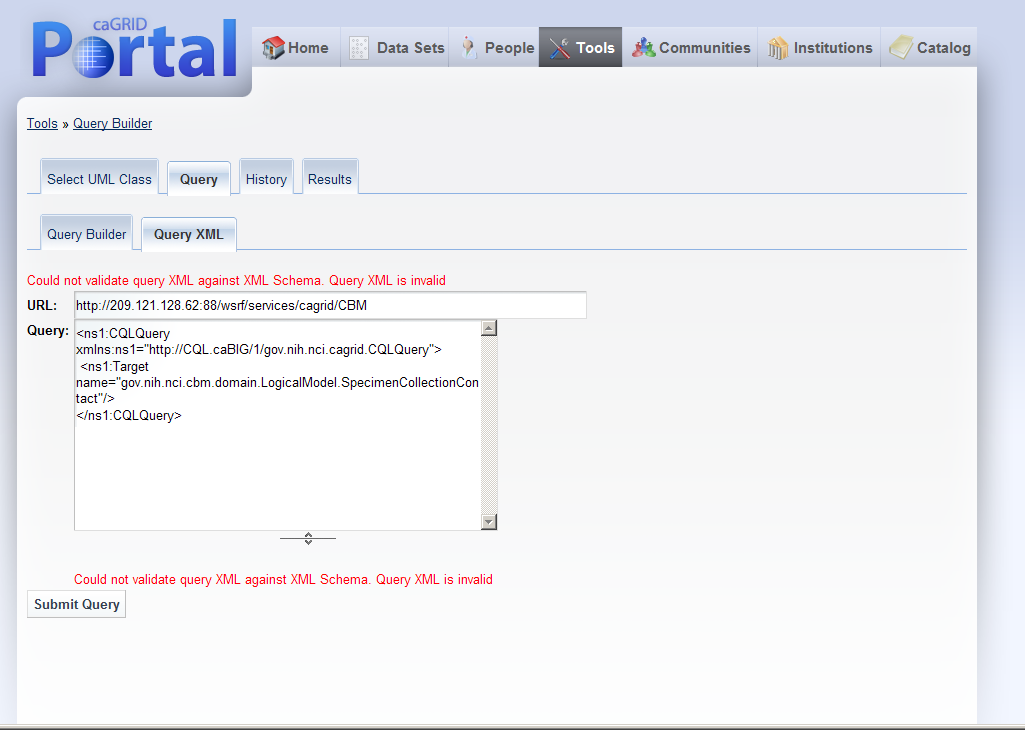




Note: for the May 2010 F2F participant, before we could enter the XML query in “Query XML” tab, and submit the query:

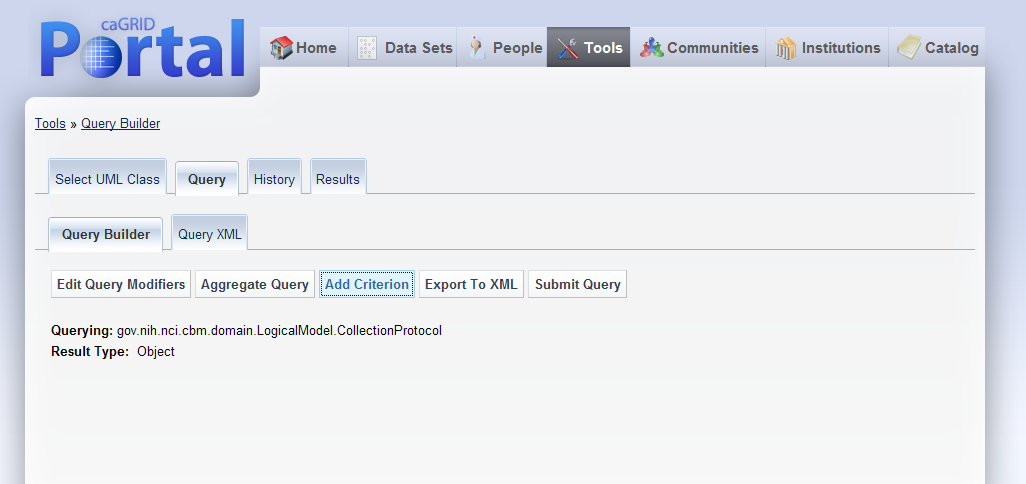


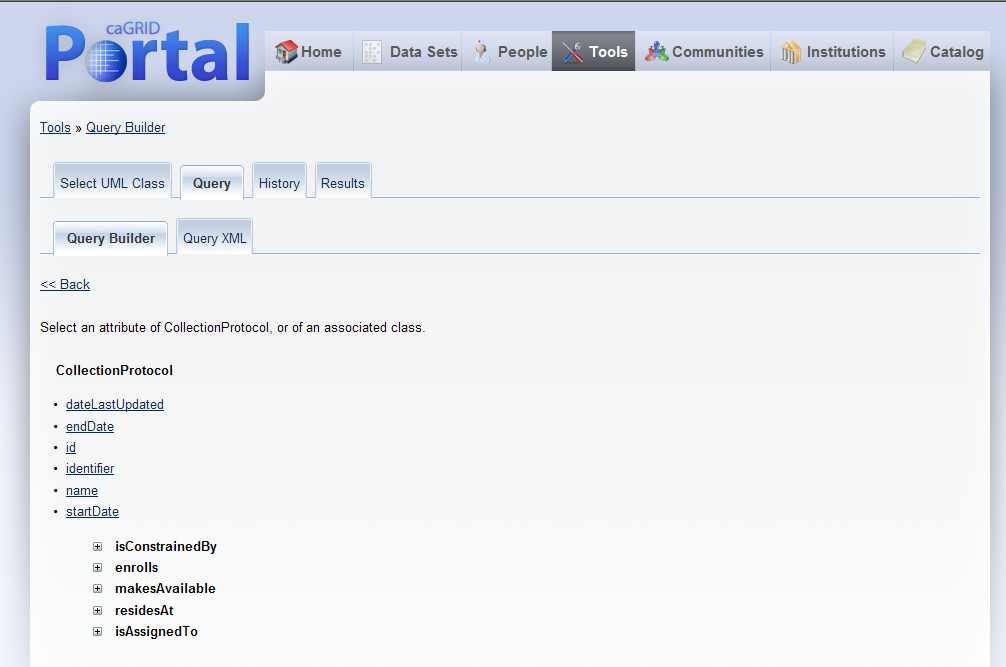
As of Oct 2010, the valid queries are producing errors such as shown below:

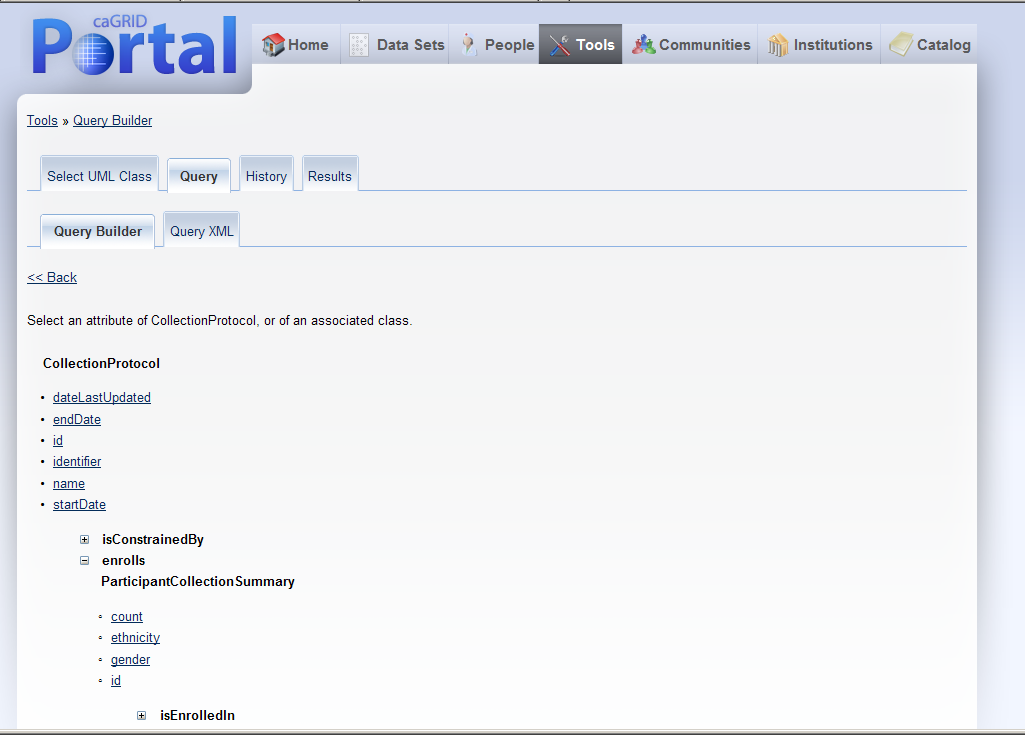


For now, it is suggested testers use the Portal for simple querying (as shown above), but use runQuery Scripts for more complex query testing.

One can also use the Grid Portal to navigate through the model, by looking at AddCriterion once a UML class is selected, and may be useful to formulate queries that should be run using the runQuery scripts







See Section 4.7, and you can test the service by replacing the URL with the appropriate URL:

ant -f queryRunnerExec.xml -Dservice.url=http://127.0.0.1:8080/wsrf/services/cagrid/CBM -Dcql.file=prostateQuery.xml

# 6.0 CBM Help Resources

|  |  |  |
| --- | --- | --- |
|  | **Contact method** | **Help provided** |
| NCI-CBM team | NCIcbm@mail.nih.gov | * Seeking to participate in CBM challenge/testing; * First stop if you have general questions on CBM goals * Inquiries on how it connects to the Specimen Resource Locator |
| CBM website | <https://cabig.nci.nih.gov/workspaces/TBPT/CBM/> | Up-to-date information and links to packages, etc. |
| TBPT –KC | <https://cabig-kc.nci.nih.gov/Biospecimen/KC/index.php/Main_Page> | * Tissue Banks and Pathology Tools Knowledge Center. |
| TBPT-KC CBM forum | <https://cabig-kc.nci.nih.gov/Biospecimen/forums/viewforum.php?f=26&sid=35545afd75b0878f9e535b424941303a> | * Discussion forum to post questions and get feedback from other CBM users, etc. * NCI-CBM team will be monitoring this forum. |
|  |  |  |
| caGrid-KC Grid-node deployment forum | <https://cabig-kc.nci.nih.gov/CaGrid/forums/viewforum.php?f=14&sid=2135e1193c95a4b4aea71987264c2980> | * Grid-KC TroubleShootingForumto post questions and discuss specific issues with Grid Deployment. |
| caGrid-KC | [knowledge@osu-citih.org](mailto:knowledge@osu-citih.org) | * Email to the caGrid-KC, but suggest to use forums first. |

## Testing Nodes Up (November 2010):

Grid Portal: <http://portal.training.cagrid.org/web/guest/home>

|  |  |
| --- | --- |
| Washington University | <http://cbm.wustl.edu/wsrf/services/caGrid/CBM> (May CBM version) |
| OSU (based on April 2010 WashU ETL | [**http://140.254.126.46:8085/wsrf/services/cagrid/CBM**](http://140.254.126.46:8085/wsrf/services/cagrid/CBM) **(May CBM version)** |
| University of Colorado | <http://ucccreg.ucdenver.edu:8080/DataService/DataService.svc> (Nov CBM version)  Data in every table; anatomic source not tracked in easily mapped way; LUNG  Diagnosis: of the 22K, only 120 had diagnosis; |
| Hollings Cancer Center | <http://128.23.35.126:8080/wsrf/services/cagrid/CBM>  (May CBM service) |

In the /cbm directory where queryRunnerExec exists:

ant -f queryRunnerExec.xml -Dservice.url=http://127.0.0.1:8080/wsrf/services/cagrid/CBM -Dcql.file=prostateQuery.xml