**The National Cancer Institute caBIG®**

**Common Biorepository Model**

**CBM Information**

**CBM 1.0 Test Grid Installation**

**December 22, 2011**

**NCI CBIIT – CBM Team**

DOCUMENT CHANGE HISTORY

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version Number | Date | Description | Author | Approved By |
| 1.0 | 05/12/11 | CBM1.0 Instructions | Anna Fernandez |  |
| 1.1 | 06/06/11 | Updated TOC | Anna Fernandez |  |
| 1.2 | 12/22/11 | Updated, direct to new service files; added description of Grid Transfer Service description (new routine in service package), and pointed to Finalized set of Grid Test validation scripts | Anna Fernandez |  |

[1.0 Common Biorepository Model Background, FAQ 4](#_Toc312407678)

[1.1 Where can I find the latest information, models, packages and test scripts for the CBM initiative? 4](#_Toc312407679)

[1.2 “CBM at a Glance” 4](#_Toc312407680)

[1.2 What does CBM 1.0 include? 7](#_Toc312407681)

[1.3 Why should CBM V1.0 (December 2011) be tested by an institute/vendor? 7](#_Toc312407682)

[1.4 How will my institute’s CBM grid service be used? 8](#_Toc312407683)

[1.5 How will having a CBM grid service at my institution help me share data and specimens? 8](#_Toc312407684)

[1.6 How will CBM be supported – responsibilities (DRAFT) 8](#_Toc312407685)

[1.7 How were the vocabulary lists generated? 9](#_Toc312407686)

[2.0 Overview CBM V1.0 Testing Grid Service Package 9](#_Toc312407687)

[2.1 What does CBM V1.0 Package.zip contain? 9](#_Toc312407688)

[3.0 Prepare the CBM Database – Extract Transform, and Load (ETL) from your Biorepository Management System database 11](#_Toc312407689)

[3.1 Select the CBM database skeleton 11](#_Toc312407690)

[3.2 ETL into CBM database 11](#_Toc312407691)

[3.2.1 Set up MySQL database 12](#_Toc312407692)

[3.2.2 Set up Oracle database 12](#_Toc312407693)

[3.2.3 Set up PostgreSQL database 12](#_Toc312407694)

[4.0 Instructions for Deploying CBM service at your test site/institute. 12](#_Toc312407695)

[4.1 Pre-requisites 12](#_Toc312407696)

[4.2 Set environment variables - CBM deployment package location 13](#_Toc312407697)

[4.3 Set up your JBoss Container 13](#_Toc312407698)

[4.4 Deploy Grid service 13](#_Toc312407699)

[4.5 Secure the JBoss Server 17](#_Toc312407700)

[4.6 Start JBoss Server 17](#_Toc312407701)

[You should see a message similar to “Hi there, this is an AXIS service!” 17](#_Toc312407702)

[5.0 CBM Marshaller – Method to provide an XML file with all CBM contents – for NCI Specimen Resource Locator use 18](#_Toc312407703)

[5.1 CBM Marshaller 18](#_Toc312407704)

[5.2 Building the CBM Marshaller 18](#_Toc312407705)

[5.3 Executing the CBM Marshaller 19](#_Toc312407706)

[6.0 Testing the CBM Grid Services that are up and running 20](#_Toc312407707)

[6.1 Using Command-Line Grid tool to check your service 20](#_Toc312407708)

[6.2 CBM Test Validation Scripts (Alpha Version – still in development) 24](#_Toc312407709)

[6.3 Looking for CBM Service over the Training caGrid Portal – 24](#_Toc312407710)

[7.0 CBM Help Resources 27](#_Toc312407711)

[APPENDIX – Installing the CBM database inside the firewall 28](#_Toc312407712)

## 

# 1.0 Common Biorepository Model Background, FAQ

## 1.1 Where can I find the latest information, models, packages and test scripts for the CBM initiative?

For a description of the goals of the CBM initiative and the latest information on the CBM Challenge, please refer to the CBM website:

<https://wiki.nci.nih.gov/display/TBPT/Common+Biorepository+Model+%28CBM%29>

## “CBM at a Glance”

In support of the U.S. National Cancer Institute’s (NCI) Best Practices for Biospecimen Resources, the Common Biorepository Model (CBM) has been developed to share up-to-date, summary-level biospecimen inventory information via the cancer Bioinformatics Grid (caBIG®) infrastructure. This enables academic and commercial researchers to identify potential partnering biorepositories from a network of reporting sites who may have appropriate specimen collections.

The CBM Service consists of:

* Thirteen pieces of information required to describe a specimen collection;
* A list of diagnoses that span cancer, cardiac, diabetic, and rare diseases in a prescribed vocabulary;
* Grid service files; and
* An example database to be used for data mapping.

Once each biorepository maps their data to the CBM and exposes de-identified information about their collections via the service, the information can be queried and updated in a periodic, automated manner controlled by each biorepository without the need for manual reporting.

More than fourteen biorepository management software vendors have participated in this initiative by reviewing early versions of the model, providing input into the classes and vocabularies, and testing early versions of the CBM services and standing up caGrid nodes on the caGrid Training Grid.

CBM service grid nodes may be queried by individual researchers via tools using the caBIG® infrastructure. One such tool will be the updated forthcoming 2011 NCI Specimen Resource Locator: an easy-to-use web interface to query and get the latest inventory information from biorepositories via CBM services.

Table . Locations of CBM-related artifacts.

|  |  |  |
| --- | --- | --- |
| Name | Location | Description |
| NCI SVN repository | <https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/>  <https://ncisvn.nci.nih.gov/WebSVN/listing.php?repname=common_biorepository_model&path=%2Ftrunk%2F> | Subversion (SVN) storage location:  Web SVN |
| \*CBM vocabulary list (Excel file): CBM\_VocabularyTerms.xls | <https://ncisvn.nci.nih.gov/WebSVN/filedetails.php?repname=common_biorepository_model&path=%2Ftrunk%2Fdatabase%2Fdata%2FCBM_VocabularyTerms.xls> | Excel file with vocabulary lists, NCI Concept Codes, and synonyms from other Standards, for use when developing ETL scripts or evaluating CBM. |
| UML of CBM model (HTML representation) | <https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/html_documentation/index.htm> | HTML view of EA model: https://ncisvn.nci.nih.gov/svn/common\_biorepository\_model/trunk/html\_documentation/index.htm Able to navigate through |
| \*UML of CBM Model (EA and .xmi formats) | <https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/caCORE_SDK/models/CBM%20with%20Value%20Domains.EAP>  <https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/caCORE_SDK/models/CBM%20with%20Value%20Domains.xmi> | Enterprise Architect Model (with Permissible values) .EA file format  and  XMI version |
| Directory for CBM schemas with populated CBM terms for use in ETL process | <https://ncisvn.nci.nih.gov/WebSVN/listing.php?repname=common_biorepository_model&path=%2Ftrunk%2Fdatabase%2F> | Each /MySQL, /Oracle, /PostgresSQL directory has:  CBM.SQL (CBM schema populated with CBM vocabulary). If available,cbm\_test\_data.zip is a test database with values for use in testing deployment. |
| \*CBM. SQL   MySQL Database | [https://ncisvn.nci.nih.gov/WebSVN/filedetails.php?repname=common\_biorepository\_model&path=%2Ftrunk%2Fdatabase%2FCBM.sql](https://ncisvn.nci.nih.gov/WebSVN/filedetails.php?repname=common_biorepository_model&path=/trunk/database/CBM.sql) | Database to be used for ETL; Now has the NCI Concept Code and NCI-Concept name |
| CBM\_1.0\_Package.zip | <https://ncisvn.nci.nih.gov/WebSVN/filedetails.php?repname=common_biorepository_model&path=%2Ftrunk%2FDeploymentPackages%2FCBM_1.0_Package.zip>  /cbm service files, CBM1.0 model,  and Grid Deployment instructions | Files publish the CBM database on caGrid; instructions, |
| \*cbmGridClient.zip | [https://ncisvn.nci.nih.gov/WebSVN/listing.php?repname=common\_biorepository\_model&path=%2Ftrunk%2FTools%2FcbmGridClientTool%2F](https://ncisvn.nci.nih.gov/WebSVN/listing.php?repname=common_biorepository_model&path=%2Ftrunk%2FTools%2FcbmGridClientTool%2F&#aea979494d01bdeb7b7fdb0ee0d30f9ef) | Command Line Run tool for performing basic querying of CBM services on the Grid (default setting: searches for CBM nodes on the Training caGrid Portal |
| CBM grid-query validation scripts | \cbm\test  \cbm\test\README file describes test | A set of test scripts that containing queries that can be run to test out the CBM service at your site.  **These are not finalized.** The final set of CBM-Grid-Tests will be used to validate the service for use by the NCI Specimen Resource Locator |
| caGrid Training Portal | https://cagrid.org/display/community/Community+Training+Grid | Current CBM testing nodes are visible. CBM service nodes should be directed to this “training caGrid Portal” site during testing/pilot phases of your implementation |
| caGrid Portal (Production) | <http://cagrid-portal.nci.nih.gov/> | caGrid Production Portal. The production CBM service should be directed to this portal to make it available for others to acess and query in production mode (by NCI Specimen Resource Locator and other locators) |

**\*Denotes items included in CBM\_1.0\_Package.zip**

CBM\_1.0\_all.zip is a test package containing:

* A revised model;
* Service files for standing on the training caGrid (for information on the caBIG Training Grid click [here](https://cagrid.org/display/community/Community+Training+Grid)) ; and
* Fixed vocabulary terms required by the CBM.

In order to display data on the caGrid, institutions must use the defined vocabulary (provided in the Excel sheet and in the model). For ease of mapping, the Excel sheet also provides a synonym mapping tab.

The Common Biorepository Model has been frozen for release 1.0. Updates to the model and accompanying vocabulary will be considered in one year. Specimen Resource Locator(SRL) Developers will take the develop documentation required for caBIG® conformance and compliance. For information on caBIG conformance click here: [Conformance Documentation](https://wiki.nci.nih.gov/display/VCDE/Introduction+to+SAIF+and+ECCF%20). The SRL Developers have begun developing the test suite that will be used for caBIG® conformance/compliance testing. As test scripts are developed they will be placed in the SVN storage location listed above.

## 1.2 What does CBM 1.0 include?

CBM V1.0 is the latest iteration of the CBM model. It has been released in order to provide sites for testing by the biospecimen management systems (vendor solutions, caTissue Suite, and in-house institute-developed solutions) so that feedback on the model, service, and testing with NCI can be provided. In this Beta version:

* The vocabularies that describe the specimen types, diagnoses types, etc. have been incorporated into the NCI Thesaurus http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources (required for caBIG® semantic interoperability) and are fixed for at least 1 year(until May 2012 or later).
* The CBM1.0 Model vocabulary has been reviewed through the caDSR (https://cabig.nci.nih.gov/concepts/caDSR/).
* 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).

The NCI would like institutes (vendors and biorepostories) to start testing and providing feedback on CBM V1.0. This version should be used to establish grid nodes for commercial off the shelf and/or custom home-grown systems

## 1.3 Why should CBM V1.0 (December 2011) be tested by an institute/vendor?

* This is the CBM v1.0 Model and service that will be used with the NCI Specimen Resource Locator (Released Dece 2011) – Information on NCI Specimen Resource Locator 2011 version can be found at: <https://wiki.nci.nih.gov/display/SRL/Specimen+Resource+Locator+wiki>
* The CBM V1.0 version has the vocabulary lists fixed for the near future the list will be fixed at a minimum through December 2012.
* Previous Beta Releases: Early limited release CBM1.0Beta 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. There are CBM nodes on the caGrid Training Grid from the CBM 1.0Beta packages.
* CBM V1.0 will be used as the backend for the Office of Biospecimens and Biorepository Research’s Specimen Resource Locator
* **NOTE:** CBM 1.0 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. The TBPT Workspace will let the community know when this happens and the validation process – but the validation tests will show that this CBM1.0 Version stood up by an institution can be used by the NCI Specimen Resource Locator

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

* Your institute’s CBM grid service will be a rudimentary (not user-friendly) service although 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 user-friendly 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 be completed in 2011 and will use CBM V1.0 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 (http://cbmi.wustl.edu/html/caB2B.html) 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?

* 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

## 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 V1.0 Testing Grid Service Package

## 2.1 What does CBM V1.0 Package.zip contain?

The package contains the files and/or links to the latest files for sites to use to ETL (Extract-Transform-Load) biorepository management system summary level information into the CBM database and deploy a caGrid CBM service associated with their site, and visible on the caGrid for querying. For biorepositories who use commercially developed specimen management systems, please talk to your solution provider to determine when this functionality will be added to your system.

**\*\*\*Unzip the CBM\_1.0\_Package.zip into a directory - this directory will be referenced as the CBM\_DIR in the instructions**

|  |  |  |
| --- | --- | --- |
| CBM\_1.0\_Package.zip Contents | Description | Location for LATEST NCI versions |
| /cbm | Build Service Files. Files will need to be modified to launch the CBM service at a biorepository | Only in package |
| /cbmGridClientTool/cbmGridClient | Command Line Run tool for performing basic querying of CBM services on the Grid (default setting: searches for CBM nodes on the Training caGrid Portal | [https://ncisvn.nci.nih.gov/WebSVN/listing.php?repname=common\_biorepository\_model&path=%2Ftrunk%2FTools%2FcbmGridClientTool%2F](https://ncisvn.nci.nih.gov/WebSVN/listing.php?repname=common_biorepository_model&path=%2Ftrunk%2FTools%2FcbmGridClientTool%2F&#aea979494d01bdeb7b7fdb0ee0d30f9ef) |
| /container/jboss-4.0.5.GA | JBoss files | Only in package |
| /database/MySQL  /database/Oracle  /database/PostgreSQL | Under each Database Type, the Directory has CBM-empty.SQL (CBM schema with not values) and CBM.SQL (CBM schema with populated values from the CBM 1.0 vocabulary lists, to be used for ETL; has the NCI Concept Code and NCI-Concept name | See Readme.txt files in each directory |
| /Database/MySQL/cbm\_test\_data.zip | MySQL CBM database contains test data for different specimen collections. For use in standing up and querying a test CBM node (before connecting with a site’s test data) | See Readme.txt files in directory |
| /Excel\_VocabList/CBM\_VocabularyTerms.xls | CBM vocabulary list: Excel file with vocabulary lists, NCI Concept Codes, definitions, and synonyms from other Standards, for use when developing ETL scripts or evaluating CBM. CBM1.0 List Finalized May 2011.  ***Excel sheet also contains useful information and guidance for ETL process (PatientAgeGroup classifications, etc.)*** | <https://ncisvn.nci.nih.gov/WebSVN/filedetails.php?repname=common_biorepository_model&path=%2Ftrunk%2Fdatabase%2Fdata%2FCBM_VocabularyTerms.xls> |
| /model | \*CBM 1.0 Model – UML  (CBM with Value Domains.EAP and .xmi formats)  Enterprise Architect Model (with Permissible values) .EAP file format  and XMI version  The EA model has the relationships and definitions of the classes, and the value domains (diagnosis, anatomic sites) definitions | <https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/caCORE_SDK/models/CBM%20with%20Value%20Domains.EAP>  <https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/caCORE_SDK/models/CBM%20with%20Value%20Domains.xmi> |
| /ws-core-4.0.3 | Directory of files for Globus toolkit (Used for Grid service deployment) | Only in package |

# 3.0 Prepare the CBM Database – Extract Transform, and Load (ETL) from your Biorepository Management System database

## 3.1 Select the CBM database skeleton

The CBM schema is provided in the following database formats

* MySQL
  + CBM service has been tested primarily against a MySQL (5.X) database.
* Oracle (pending)
* PostgreSQL (pending)

Note: Some early CBM beta testers have also used Microsoft SQL (University of Colorado). Please contact TBPT for more information.

## 3.2 ETL into CBM database

There are 3 files useful for reviewing when defining the ETL scripts:

* The **CBM 1.0 EA model** (*CBM with Value Domains.EAP*, *CBM with Value Domains.xmi*)
  + - 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](https://ncisvn.nci.nih.gov/svn/common_biorepository_model/trunk/html_documentation/index.htm):
* CBM\_VocabularyTerms.xls
  + Please refer to the */Excel\_VocabList* directory for the Excel file that contains mappings to SNOMED, ICD, and synonyms found in NCI Thesaurus for the terms used in CBM1.0Beta
  + Vocabulary lists –all now included in the NCI MetaThesaurus (required for semantic/syntactic compatibility)
* *CBM.SQL (available under each Database Directory in /database/\*)*
  + This database is CBM schema, “populated” with the diagnosis list of values and the other terms required for reference.
  + This SQL schema can be used for ETL from the CBM-institute’s biorepository system. This schema has the expected domain values for each class within the model.
* Pentaho Kettle or other preferred ETL programs can be used. Some working versions of Kettle scripts for mapping into caTissue Suite (not guaranteed to work) are found [here](https://ncisvn.nci.nih.gov/WebSVN/listing.php?repname=common_biorepository_model&path=%2Ftrunk%2FETL%2F&#a2cda98207395d94a970ddede5957e3be).
* It is recommended to use the CBM.SQL along with the CBM\_VocabularyTerms.xls file during your mapping process.
* Save your CBM database into a preferred database name: e.g. cbm.sql

**For Testing of CBM service:** There is a **cbm\_test\_data.sql** with valid queries in the /database directory (available for MySQL).

### 3.2.1 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. The database must be configured to treat table names in a case insensitive way. This is the default on Windows and other environments that have case-insensitive file names. On Linux, MAC-OS and other UNIX-like environments, insert the following line in the file /etc/my.cnf after the line the contains [mysqld]:  
   lower\_case\_table\_names=1
5. Create a user that has permissions to access the database: <http://dev.mysql.com/doc/refman/5.1/en/adding-users.html>
6. Load the test data into the mysql database
   1. Example using mysql command prompt: source cbm\_test\_data.sql

Note: A simple database (cbm\_test\_data.sql) is included with limited data. One can use CBM.SQL to create the site’s individual cbm\*.sql database and populate sample data using MYSQL programs.

### 3.2.2 Set up Oracle database

TBD

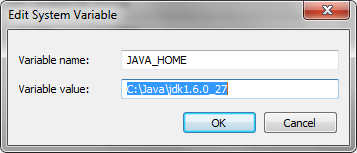
### 3.2.3 Set up PostgreSQL database

TBD

# 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 version of JDK 5 (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.



(Note: To avoid some potential errors, better to have the directories NOT have spaces, as shown above).

1. 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 Set environment variables - CBM deployment package location

CBM\_DIR will refer to the name of the directory where you unzipped the package (this directory should have /cbm, etc.)

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

## 4.3 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>

1. Verify the valid JDBC drivers are available in the JBoss directory JBOSS\_HOME/server/default/lib
   1. **MySQL.** MySQL JDBC jar drivers for MySQL are included in the directory.
   2. **Oracle and PostgreSQL**. Check the JBOSS\_HOME/server/default/lib directory for a copy of the appropriate JDBC driver jar for your database. Copy your JDBC jar there if it is missing.
      1. The proper Oracle and Postgres JDBC drivers (jars) will need to be installed in this directory and will be dependent on the database versions (distributions) used.

## 4.4 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 (Working copy location)
3. Update file in **cbm-orm.jar (Direct to MySQL, Oracle (pending), and PostgreSQL (pending) database type)**
   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:**

Follow the commented fields in the file that denote MySQL, Oracle, and PostGre settings. Below is a description of the items that are changing.

* + 1. Update the database connection information: (in the test package, “cbmTestData” based off cbm\_test\_data.SQL schema can be used if the ETL process in your site is incomplete.

Below, highlighted areas denote the values that should be specific to your install, example shows “cbm” as the name of the site database)

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

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

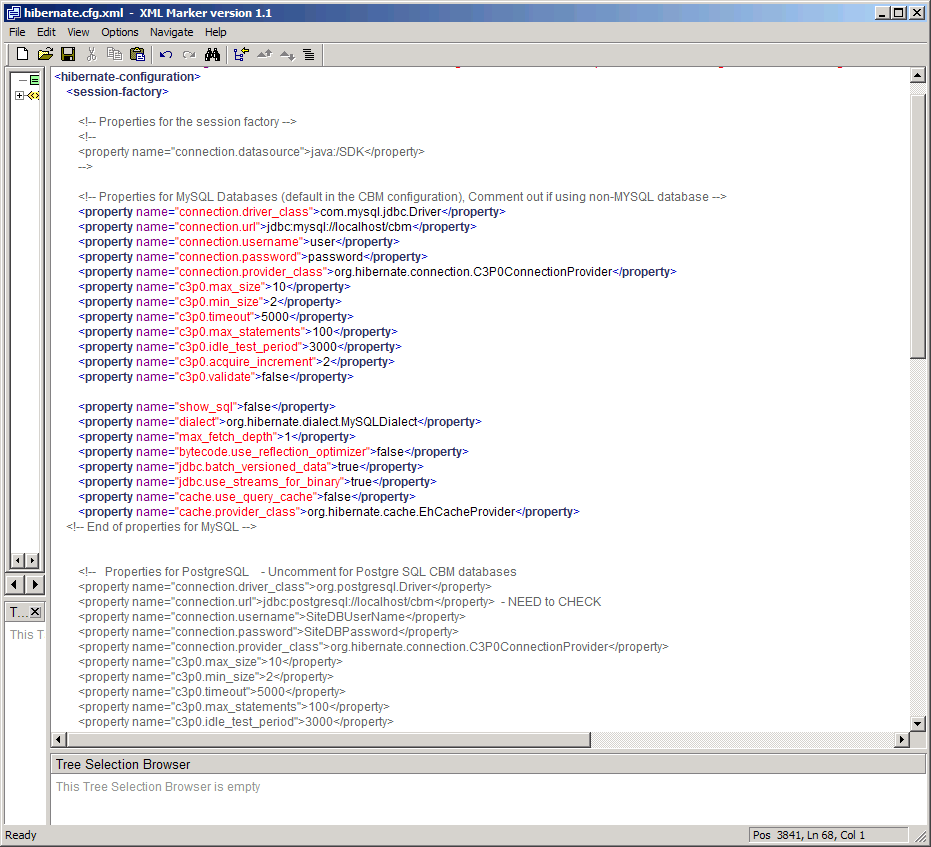
<property name="connection.password">password</property>

* + 1. Update the connection.driver\_class property to list the driver name of the database server being used.

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

* + 1. Update the Hibernate dialect property to align with the database server being used.

<property name="dialect">org.hibernate.dialect.MySQLDialect</property>



**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.

Re-create/update the jar with the new file (i.e. in the CBM-orm directory, run the command: …/jdk1.5.0\_22\bin\jar –cvf CBM-orm.jar \*)

1. Copy the updated **CBM-orm.jar** to the cbm/lib directory
2. Edit **cbm/etc/serviceMetadata.xml** to associate with your site’s contact information, for use in caGrid settings.

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="XXX" postalCode="" stateProvince="OH" street1="" street2=""/>

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

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

**NOTE:** make sure the site completes the contact information (name, email, address) for all “PointOfContact” and “hostingResearchCenter” information.

1. Update the index service registration URL

**Note:** default is set to the Training caGrid. As sites are testing, set the service to point to the Training grid. Once the ETL and data is associated with real biorepository content, direct the service to the Production grid.

* 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: During testing, 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***

(NOTE: ant undeployJBoss; ant deployJBoss –logfile XXX.txt to record output)

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 query program already knows  the  hostname or ip address, they can call the  service  directly.

## 4.5 Secure the JBoss Server

To secure the JBoss container, please follow one of the options detailed in this wiki site (download and run the secureJBoss script, or delete files)

<https://wiki.nci.nih.gov/display/BuildandDeploymentAutomation/Securing+JBOSS+Console+Apps>

## 4.6 Start JBoss Server

1. Run the command: **$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>

## You should see a message similar to “Hi there, this is an AXIS service!”

# 5.0 CBM Marshaller – Method to provide an XML file with all CBM contents – for NCI Specimen Resource Locator use

https://wiki.nci.nih.gov/display/SRL/Creating+an+XML+dump+file+from+a+CBM+database+%281c%29 Contains the latest information.

## 5.1 CBM Marshaller

All data that is imported into the Specimen Resource Locator comes in the form of a CBM-formatted XML file.  This CBM file may be delivered to the SRL either by automated downloads from CBM nodes, or by manual distribution of the XML files to the SRL administrator.  The CBM structured XML files can be generated from a CBM database by using the CBM Marshaller Java application.  The CBM Marshaller is Java application that is included in the CBM distribution.  The purpose of the CBM Marshaller is to extract data from a CBM-structured database and save the data to an XML file that conforms to the CBM.xsd XML schema.  These properly formatted XML files can then be loaded into the NCI SRL.

The files that are generated by the CBM Marshaller can be provided to the SRL either by emailing the files to the SRL administrator, or by establishing a CBM node

## 5.2 Building the CBM Marshaller

Prior to executing the CBM Marshaller it must be built.

The CBM Marshaller build file is named marshaller-build.xml and is located in the root directory of the CBMService project directory.  The CBM Marshaller build target requires several arguments.  These arguments are used to populate the database connection information in the fully built CBM Marshaller application.

The arguments required for the CBM Marshaller build target are:

|  |  |
| --- | --- |
| Name | Description |
| db\_driver | The name of the JDBC Driver to be used by the CBM Marshaller.  Common values for these drivers are:   **MySQL**: com.mysql.jdbc.Driver  **PostgreSQL**: org.postgresql.Driver  **Oracle**: oracle.jdbc.driver.OracleDriver   Please note:  The CBM distribution includes MySQL 5.x and PostgresSQL 8.3 drivers.  Copy additional JDBC driver jars to the CBMService/lib directory as needed |
| db\_protocol | The protocol section of the JDBC connection string.  Common values for protocol name are:   **MySQL**: jdbc:mysql  **PostgreSQL**: jdbc:postgresql  **Oracle**: jdbc:oracle:thin |
| db\_host | The name of the CBM database’s host server |
| db\_port | The port number of the CBM database instance |
| db\_name | The name of the CBM database |
| db\_user | The user name required to access the CBM database |
| db\_password | The password required to access the CBM database |
| db\_dialect | The Hibernate dialect of the database.  Common values for dialect name are:   **MySQL**: org.hibernate.dialect.MySQLDialect  **PostgreSQL**: org.hibernate.dialect.PostgreSQLDialect  **Oracle**: org.hibernate.dialect.OracleDialect |

 Sample execution command for the CBM Marshaller build target:

|  |
| --- |
| ant -buildfile marshaller-build.xml build -Ddb\_driver=com.mysql.jdbc.Driver -Ddb\_protocol=jdbc:mysql -Ddb\_host=localhost -Ddb\_port=3306 -Ddb\_name=cbm -Ddb\_user=root -Ddb\_password=root -Ddb\_dialect=org.hibernate.dialect.MySQLDialect |

 Executing the CBM Marshaller build target will create a new package directory in the CBM Service directory structure.  The package directory will contain:

* The CBM Marshaller executable jar
* The CBM Marshaller Hibernate configuration file
* The CBM.xsd XML schema file
* A lib directory containing all library jars required for execution of the CBM Marshaller

## 5.3 Executing the CBM Marshaller

Once the CBM Marshaller build is completed, it is ready for execution.  The CBM Marshaller works by extracting data from a CBM formatted database using the set of CBM Hibernate configuration files.  All data from the CBM data is extracted, and written to an XML file which conforms to the CBM.xsd XML schema definition.

The CBM Marshaller execute target requires two arguments.  These arguments are:

|  |  |
| --- | --- |
| Name | Description |
| file | The name of the CBM data extract file to be written |
| info | Source information of the CBM data.  ex) WashU caTissue Data |

Sample execution command for the CBM Marshaller execute target:

[?](https://wiki.nci.nih.gov/display/SRL/Creating+an+XML+dump+file+from+a+CBM+database+%281c%29)

|  |
| --- |
| ant -buildfile marshaller-build.xml execute -Dfile=./cbm.xml -Dinfo="My CBM Data" |

# 6.0 Testing the CBM Grid Services that are up and running

* 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: In the package, there is a set of Test Scripts in /cbm/test with a Readme file**
* Below are 3 methods to test: 1) Command line tool for testing and querying the Grid services that are published on the training grid (can be used for debugging the CBM service to assess where the problems may lie), and 2) Alpha-version of the CBM grid test suite that will be used for CBM service validation(to be used by sites and the NCI Specimen Resource Locator), and 3)GridPortal viewing and basic querying

## 6.1 Using Command-Line Grid tool to check your service

The caGrid-Knowledge Center Team has developed a command line Grid Tool. For in-depth information on the Grid Client, please see the caGrid Client Application Guide: <https://cagrid.org/display/knowledgebase/Client+Application+Guide>

There is a customized version of the tool for CBM data services provided in the CBM deployment package.

**Install:**

* In a command window, go to **CBM\_DIR/cbmGridClientTool/,** there should be /cbmGridClient directory there
* From a command line in the /cbmGridClient directory, run the 'ant all' command.

**Run:**

* From a command line, run the 'ant run' command

**Setting to use Production caGrid (use this once your CBM site is ready to go to production, default tool is set to look at the caGrid Training Grid):**

1. Open the file cbmGridClient/ivy.xml
2. Search for the text 'Target Grid Configuration'
3. To configure for production, change 'training-1.3' to 'nci\_prod-1.3'
4. Save the file
5. From a command line, run the 'ant clean all' command

**Run Options:**

When the client is run you will be presented with 5 options:

1. **Sync With Trust Fabric**

This option will run the SyncGTS client.  This retrieves the grid certificates into the .globus directory.

Example Output:

    [java] Select client action:  [1..5] :

1

     [java]

     [java] Synchronize Once...

     [java] Sync file = conf/sync-description.xml

     [java] A pre synchronization snapshot of the Trusted CA Directory found 8 Trusted CAs.

     [java] The CA O=caBIG,OU=caGrid,OU=Training Trust Fabric,CN=caGrid Training Trust Fabric CA was not removed because it is the exclude list.

     [java] The CA O=caBIG,OU=caGrid,OU=Training Trust Fabric,CN=caGrid Training Trust Fabric CA was not removed because it is the exclude list.

     [java] Successfully removed 6 Trusted Authority(s) from /Users/williamstephens/.globus/certificates

     [java] Successfully wrote out 6 Trusted Authority(s) to /Users/williamstephens/.globus/certificates

     [java] Synchronize Complete.

     [java]

2. **Login to Grid**

This option allows a user to authenticate against the Dorian service using their grid username and password.

A grid credential will be written to cbmGridClient/user.proxy.  Subsequent calls to login will use this file while it is valid, rather than request username and password again.

Example Output:

     [java] --------------------------

     [java] 1 : Sync with Trust Fabric

     [java] 2 : Login to Grid

     [java] 3 : Search for CBM Services in Index Service

     [java] 4 : Query a CBM Data Service

     [java] 5: Quit

     [java] Select client action:  [1..5] :

2

     [java]

     [java] Login to the Grid

     [java] Provide Username:

     [java]

wistephens

     [java] Provide Password:

     [java]

<Password will be shown>

     [java] Enter credential lifetime in hours [1..12] :

12

     [java] Identity = /O=caBIG/OU=caGrid/OU=Training/OU=Dorian/CN=wistephens

     [java] User credential saved to  = user.proxy

     [java] Login complete.

3. **Search for CBM Services in Index Service**

This option allows the user to query the Index service for all instances of CBM services.

Example Output:

     [java] Select client action:  [1..5] :

3

     [java]

     [java] Discovering Grid Services

     [java] Searching for 'CBM' services

     [java] Available services:

     [java]     0: <http://209.121.128.62:88/wsrf/services/cagrid/CBM>

     [java]     1: <https://cabigapps01.cancer.gov:9295/wsrf/services/cagrid/CBM>

     [java]     2: <http://tutorials.training.cagrid.org:8085/wsrf/services/cagrid/CBM>

     [java]     3: <http://128.23.35.126:8080/wsrf/services/cagrid/CBM>

**4. Query a CBM Data Service**

When specifying that you want to "Query a CBM Data Service" you will be provided with the option of selecting the service from the Index Service or providing a URL (e.g. if you are testing on <http://localhost:8080/wsrf/CBM>, etc) .

     [java] --------------------------

     [java] 1 : Sync with Trust Fabric

     [java] 2 : Login to Grid

     [java] 3 : Search for CBM Services in Index Service

     [java] 4 : Query a CBM Data Service

     [java] 5: Quit

     [java] Select client action:  [1..5] :

4

     [java]

     [java] Query Data Services:

     [java] 1 : Select CBM service from the Index Service

     [java] 2 : Provide Local CBM service URL

     [java] Select service location [1..2] :

2

     [java] Provide the URL of the local service:

<http://tutorials.training.cagrid.org:8085/wsrf/services/cagrid/CBM>

     [java] 1 : Select query object from list

     [java] 2 : Read CQL query from file

     [java] Select Query Type [1..2] :

1

This option allows the user to select a CBM service and submit a query to the service using 2 options:

1. Select an object that is exposed via the CBM domain Model
2. Specify the name of a CQL query file that is stored in the cbmGridClient directory, with the query output written to the file 'cbmGridClient/queryResults.xml'

Example Output 1:

     [java] Select client action:  [1..5] :

4

     [java]

     [java] Query Data Services: obtaining service list...

     [java] Searching for 'CBM' services

     [java] Available CBM Data Services:

     [java]     0: <http://209.121.128.62:88/wsrf/services/cagrid/CBM>

     [java]     1: <https://cabigapps01.cancer.gov:9295/wsrf/services/cagrid/CBM>

     [java]     2: <http://tutorials.training.cagrid.org:8085/wsrf/services/cagrid/CBM>

     [java]     3: <http://128.23.35.126:8080/wsrf/services/cagrid/CBM>

     [java] Select Service [0..3] :

2

     [java] 1 : Select query object from list

     [java] 2 : Read CQL query from file

     [java] Select Query Type [1..2] :

1

     [java]     0: gov.nih.nci.cbm.domain.LogicalModel.Organization

     [java]     1: gov.nih.nci.cbm.domain.LogicalModel.Institution

     [java]     2: gov.nih.nci.cbm.domain.LogicalModel.Person

     [java]     3: gov.nih.nci.cbm.domain.LogicalModel.CollectionProtocol

     [java]     4: gov.nih.nci.cbm.domain.LogicalModel.SpecimenCollectionSummary

     [java]     5: gov.nih.nci.cbm.domain.LogicalModel.Address

     [java]     6: gov.nih.nci.cbm.domain.LogicalModel.SpecimenCollectionContact

     [java]     7: gov.nih.nci.cbm.domain.LogicalModel.AnnotationAvailabilityProfile

     [java]     8: gov.nih.nci.cbm.domain.LogicalModel.ParticipantCollectionSummary

     [java]     9: gov.nih.nci.cbm.domain.LogicalModel.Race

     [java]     10: gov.nih.nci.cbm.domain.LogicalModel.Preservation

     [java]     11: gov.nih.nci.cbm.domain.LogicalModel.Diagnosis

     [java]     12: gov.nih.nci.cbm.domain.LogicalModel.SpecimenAvailabilitySummaryProfile

     [java] Select a object to query [0..12] :

9

     [java] Querying: gov.nih.nci.cbm.domain.LogicalModel.Race

     [java] <ns2:Race race="Black or African American" id="602" xmlns:ns2="gme://caCORE.caCORE/3.2/gov.nih.nci.cbm.domain.LogicalModel"/>

     [java] <ns3:Race race="Not Reported" id="603" xmlns:ns3="gme://caCORE.caCORE/3.2/gov.nih.nci.cbm.domain.LogicalModel"/>

     [java] <ns4:Race race="Unknown" id="604" xmlns:ns4="gme://caCORE.caCORE/3.2/gov.nih.nci.cbm.domain.LogicalModel"/>

     [java] <ns5:Race race="Asian" id="622" xmlns:ns5="gme://caCORE.caCORE/3.2/gov.nih.nci.cbm.domain.LogicalModel"/>

     [java] <ns6:Race race="American Indian or Alaska Native" id="647" xmlns:ns6="gme://caCORE.caCORE/3.2/gov.nih.nci.cbm.domain.LogicalModel"/>

     [java] <ns7:Race race="White" id="1475" xmlns:ns7="gme://caCORE.caCORE/3.2/gov.nih.nci.cbm.domain.LogicalModel"/>

     [java] <ns8:Race race="Native Hawaiian or Other Pacific Islander" id="1496" xmlns:ns8="gme://caCORE.caCORE/3.2/gov.nih.nci.cbm.domain.LogicalModel"/>

     [java]

Example Output 2:

     [java] Select client action:  [1..5] :

4

     [java]

     [java] Query Data Services: obtaining service list...

     [java] Searching for 'CBM' services

     [java] Available CBM Data Services:

     [java]     0: <http://209.121.128.62:88/wsrf/services/cagrid/CBM>

     [java]     1: <https://cabigapps01.cancer.gov:9295/wsrf/services/cagrid/CBM>

     [java]     2: <http://tutorials.training.cagrid.org:8085/wsrf/services/cagrid/CBM>

     [java]     3: <http://128.23.35.126:8080/wsrf/services/cagrid/CBM>

     [java] Select Service [0..3] :

2

     [java] 1 : Select query object from list

     [java] 2 : Read CQL query from file

     [java] Select Query Type [1..2] :

2

     [java] Provide CQL Filename:

cbmQuery.xml

     [java] Querying non-secure with CQL query

     [java] Query results written to queryResults.xml

     [java]

## 6.2 CBM Test Validation Scripts (Finalized, in /cbm)

NCI is creating a set of test-scripts that will validate the CBM service at a site is valid and is able to be used to extract out useful information from the CBM service.

The test scripts and instructions to run these CBM grid-query validation scripts are found in the

/cbm/test Directory, with associated README file explaining how to run them

Also, this wiki page describes how they are used with respect to validation for use with the NCI Specimen Resource Locator:

https://wiki.nci.nih.gov/display/SRL/Data+validation+and+publish+%28step+3%29

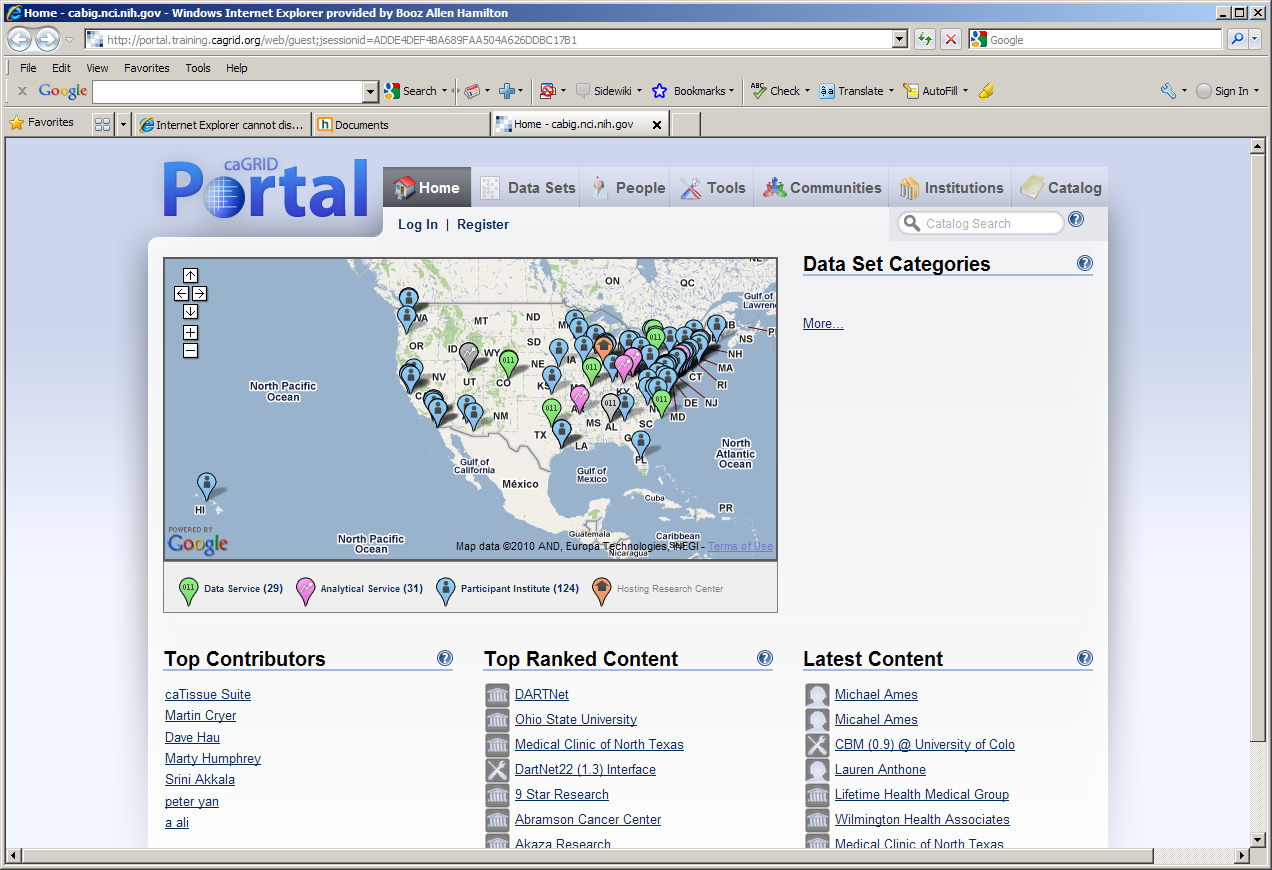
Note: A set of test scripts that containing queries that can be run to test out the CBM service at your site.

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

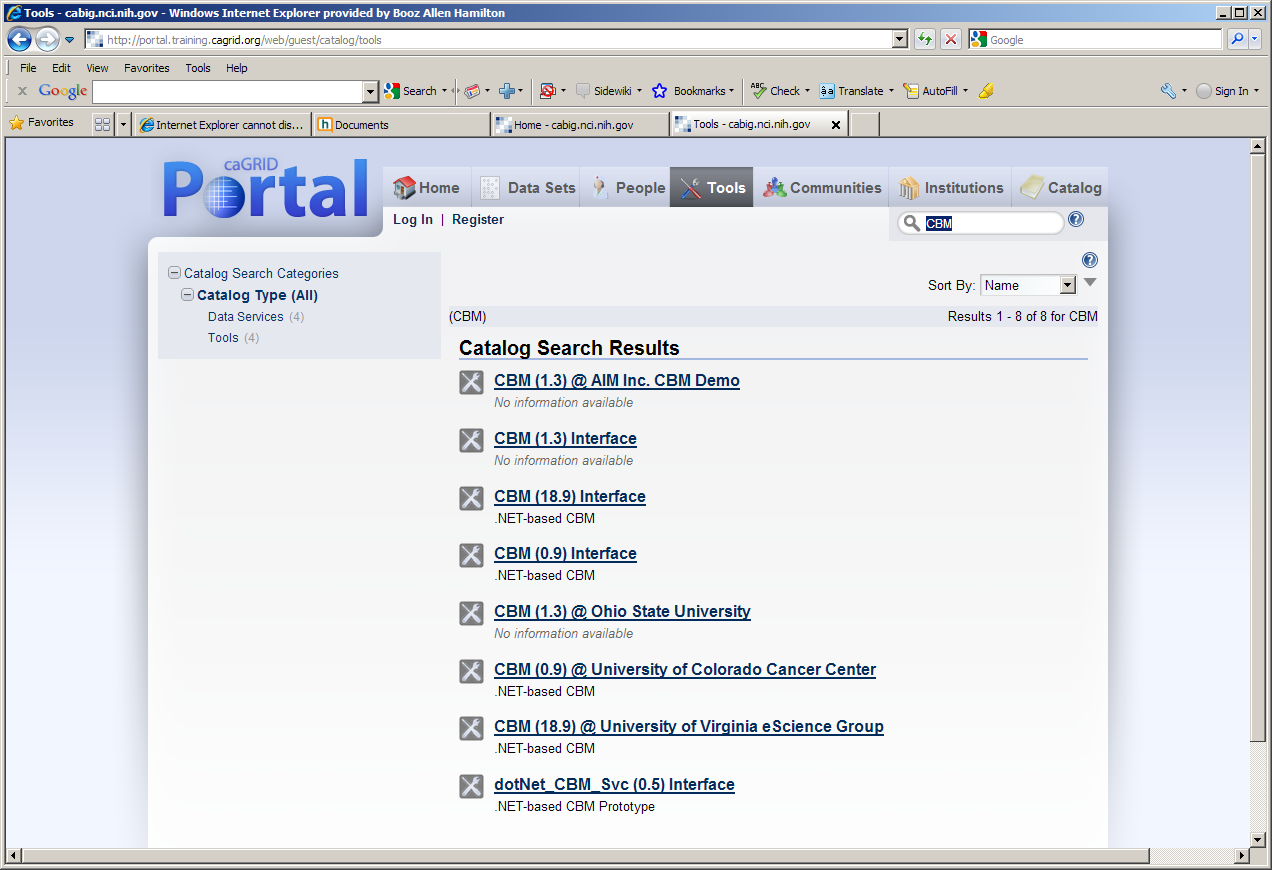
* You will be able to query the grid on the Training Grid portal:

How to find the grid service:

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

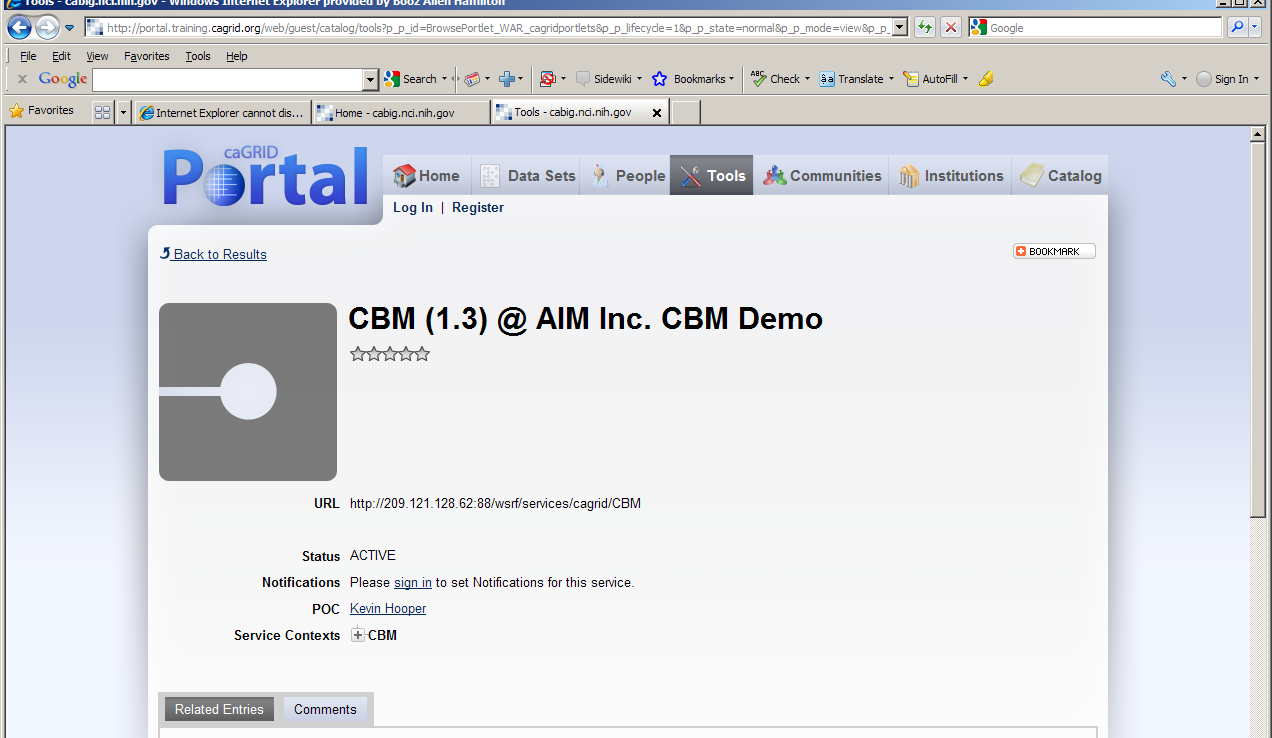


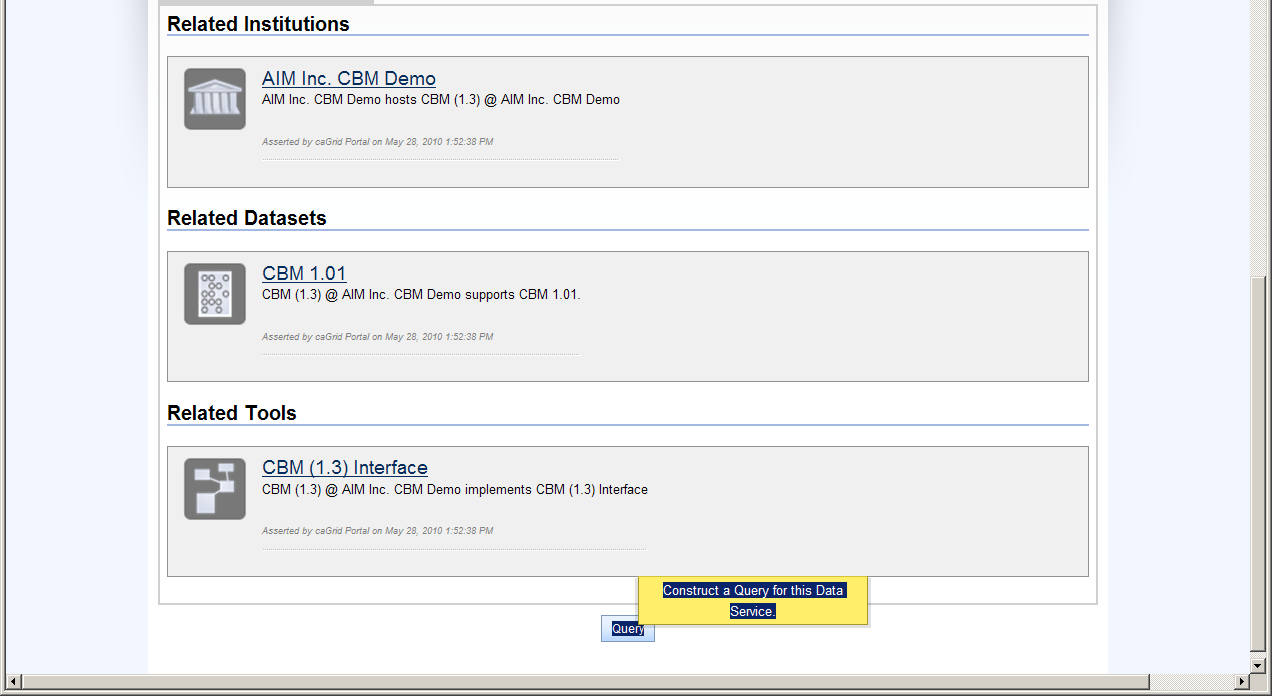
Go to top bar and click on “Tools” and then in the “Catalog Search”, enter “CBM”



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

Clicking on the service link:





# 7.0 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. |

## APPENDIX – Installing the CBM database inside the firewall

The parameters you ask about would go in   
$JBOSS\_HOME/server/default/deploy/wsrf.war/WEB-INF/etc/globus\_wsrf\_core/server-config.wsdd

The parameters may be needed to get registration with the indexing service to work correctly when the service is behind a firewall and communicates with the outside world through a proxy. These should not be needed with the service running in a DMZ, since it is accessible from the public internet without having to go through a proxy.

As we have explained, getting caGrid services to work through a proxy can pose a variety of problems.  The particular problem that these two parameters solve is a caGrid service publishing an address that is not accessible from outside its firewall to an index service.  When this happens, clients operating outside the firewall can see the service’s advertisement in the index, but are unable to communicate with the service through the published address.

The solution to this problem is to publish the address of the proxy, rather than the service’s own address, to the index.

*<parameter* *name="logicalHost"* *value="128.23.35.126"/>*

Tells the service to publish the address *128.23.35.126* instead of its own address. If 128.23.35.126 is the address of the service’s proxy, then that solves the immediate problem but gives rise to another.

By default, a caGrid service will not publish its IP address, but its DNS address instead.  If the proxy has a DNS address *and* the proxy’s DNS address is published to public DNS servers, then this is not a problem.  If the proxy has a DNS address that is not published to public DNS servers, then publishing the proxy’s DNS address will not be helpful since the DNS address will not be resolvable outside the firewall.  To solve this problem we specify the parameter

*<parameter* *name="publishHostName"* *value="false"/>*

This tells the service to publish the IP address instead of the DNS address.