Contract No. HHSN272001200010C

Immune Epitope Database and Analysis Program

System Architecture and Database Design Specification, v3.1
Curation v2.9.10 and External v3.6.0 releases

La Jolla Institute for Allergy and Immunology
9420 Athena Circle
La Jolla, CA 92037

858-752-6923
858-752-6987 (fax)
wfleri@LJI.org

December 1, 2016

Table of Contents

List of Tables

List of Figures

IEDB System Architecture and Database Design Revision History

Revision	Date	Description
1.0	August 23, 2010	Initial Release
2.0	September 30, 2012	First post-renewal release
3.0	March 31, 2015	First release of IEDB 3.0
3.1	March 31, 2017	Update after adding the polling server

1.0 Introduction

1.1 Scope

The scope of this document is to provide an architectural overview of the Immune Epitope Database (IEDB) system hardware, database design and available applications. In doing so this document will capture and convey the infrastructure and high-level database design which forms the basis for the key IEDB system components, the Curation application and the External or public facing application.

1.2 Purpose

The purpose of the IEDB system is to provide the scientific community a free public repository of immune epitope data. The web-based IEDB system surpasses previously available immune epitope databases by including detailed description of the experimental and immunological context in which epitopes are recognized. The IEDB contains curated data relating to all infectious diseases, including category A-C pathogens, emerging and re-emerging infections diseases, allergens, diabetes, rheumatoid arthritis, multiple sclerosis and lupus as of December 2009. A team of highly trained curators continue to add more immune epitope data to the IEDB on a weekly basis. The curators read scientific journal articles, identify immune epitope details and associated data and enter data into the IEDB. A Data Submission Tool (DST) is also integrated with the IEDB to facilitate external submission of data that supplement or is independent of published journal articles.

The public facing IEDB application provides extensive query capabilities against the underlying database. The IEDB Analysis Resource provides a collection of specialized tools for more complex data analysis and prediction of epitopes. The target audience for the IEDB includes any person capable of accessing the Internet; however the primary users are those within the scientific community. The IEDB staff has worked diligently to successfully create a close working relationship with the scientific community. User feedback and help requests are used to continuously improve functionality and add new features to the IEDB system.

1.3 Assumptions and Dependencies

The following is a list of identified assumptions and dependencies for the IEDB project:

Assumptions:

- This system will be constructed as a web site.
- The features intended for the public will be accessible on the Internet.
- The system will be accessed using an Internet browser.
- The system will not be constructed to address classified data.
- System security will be based on permissions granted to authenticated users.

Dependencies:

- The system will utilize taxonomy data from NCBI.
- The system will interact with PDB using web services.
- The system will interact with GenBank using web services.
- The system will interact with GenPept using web services.
- The system will retrieve citations from PubMed.
- The system will retrieve information from ChEBI web services.
- Add others?

1.4 Key Objectives

The key objective of the IEDB System Architecture and Database Design Document is to provide a high-level overview of the system architecture and data models to support the database design. This document will be revised to support new IEDB design features or any modifications. Revisions to this document will follow the change procedure defined below in section 1.7.

1.5 Contract Identification

La Jolla Institute for Allergy and Immunology (LJI) has developed, managed and maintained the IEDB since December 2003 under contract number HHSN266200400006C and a seven-year renewal contract number HHSN272001200010C. The following subcontractors assist LJI in the ongoing development and maintenance of the project: Leidos and the Technical University of Denmark (DTU).

1.6 References

Table 1-1. IEDB Project, Technical and Scientific Resources

IEDB Project Documents
RFP No. NIH-NIAID-DAIT-03-31
IEDB Curation Schema, Version 2.4
IEDB Curation Manual, September 02, 2009
IEDB User Documentation Version 2, January 21, 2009
IEDB Annual Compendium for2015, May 6, 2016
IEDB LinkOut procedure, July 22,2009
Field Guide for DST, Beta, April 15, 2009
The Ontology of Immune Epitopes (ONTIE) at http://ontology.immuneepitope.org/
Technical Materials
Designing Enterprise Applications with the J2EE Platform, 2nd Edition, Sun Microsystems, 2004

Mastering BEA WebLogic Server: Best Practices for Building and Deploying J2EE Applications by Gregory Nyberg, Robert Patrick et al. John Wiley & Sons, 2003
Struts: The Complete Reference by James Holmes, McGraw-Hill/Osborne, 2004
Mastering Jakarta Struts by James Goodwill, John Wiley & Sons, 2002
Professional Java Server Programming by Danny Ayers, Wrox Press Ltd., 1999
JUnit in Action by Vincent Massol, Manning Publications Co., 2004
Professional XML by Diedier Martin, Wrox Press Ltd., 1999
Instant UML by Pierre-Alain Muller, Wrox Press Ltd., 2000
Apache Lucene at http://lucene.apache.org/java/docs/index.html
PHP at www.php.net
Spring Framework at http://www.springsource.org/
Display Tag Library at http://displaytag.sourceforge.net/1.2/index.html
JSON at http://json.org/
Ajax: The Complete Reference, by Thomas Powell, McGraw-Hill Osborne Media 2008
MySQL database documentation at www.mysql.com/doc/refman/5.1/en/index.html
JavaScript: The Complete Reference, 2nd Edition, by Thomas Powell, McGraw-Hill Osborne Media, 2004
Javascript, CSS, PHP reference material at www.w3schools.com
Apache Web Server documentation at httpd.apache.org/docs/2.2/
Dojo JavaScript library documentation at www.dojotoolkit.org/reference-guide/
Apache Tomcat documentation at tomcat.apache.org/tomcat-5.5-doc/index.html
Scientific Materials
Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, Peters B. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2014 Oct 9. pii: gku938. [Epub ahead of print] PubMed PMID: 25300482.
The Ontology for Biomedical Investigations (OBI) at http://ontology.immuneepitope.org/

1.7 Change Procedures

This document is under change control and will continue to be updated as necessary. The revision history of this document is tracked in the Revision History section at the beginning of this document, page iii. Significant changes to the document will be represented by the new version as being incremented by the next whole number (e.g. 2.0). Minor updates to this document will be traced as minor releases (e.g.1.1).

2.0 System Architecture

2.1 System Overview

The IEDB system is comprised of three main components: Curation (in which data is entered into the database), External (which is accessed by external users to query the database), and Document Management (which serves to identify relevant references and track their curation). The IEDB also interfaces with different external resources to obtain data necessary for detailed curation of a reference. Figure 2-1 below provides a high level representation of the IEDB system design.

2.1.1 Document Management Overview

The Document Management component of the IEDB consists of the Curation Tracking System, the Document Classifier and the Document Retriever. Document Management identifies what journal articles should be curated, downloads a copy of the publication and tracks the status of the article curation. The first step in curation is identifying relevant publications and is illustrated in Figure 2-2. The Document Classifier uses queries of PubMed and the Protein Data Bank (PDB), along with machine learning methods to identify all potential curation references. It is implemented as a MySQL database and a set of Python scripts. Once a journal article has been classified as a curation item, the Document Retriever fetches and downloads the associated PDF file and makes it available for a curator to pick in the curation component of the system. The Curation Tracking System (CTS) keeps track of each reference from when it is assigned to a curator and all of the way through the curation process until it is approved and promoted to production.

2.1.2 IEDB Curation Overview

The IEDB Curation toolset is only accessible by curation staff. This internal web-based application includes a curation toolset that is used to perform all the "behind the scene" functions such as curation, internal user administration, and system configuration. In addition, Curation includes a Data Submission Tool (DST). The DST allows any user the ability to submit data directly to the IEDB for curation. Submitted data are then transferred to the curation toolset, where the submission is manually reviewed by curation staff. Once epitope data has been entered, reviewed, and approved, it is released to IEDB External for public consumption. Section 3.6 contains the Curation data models.

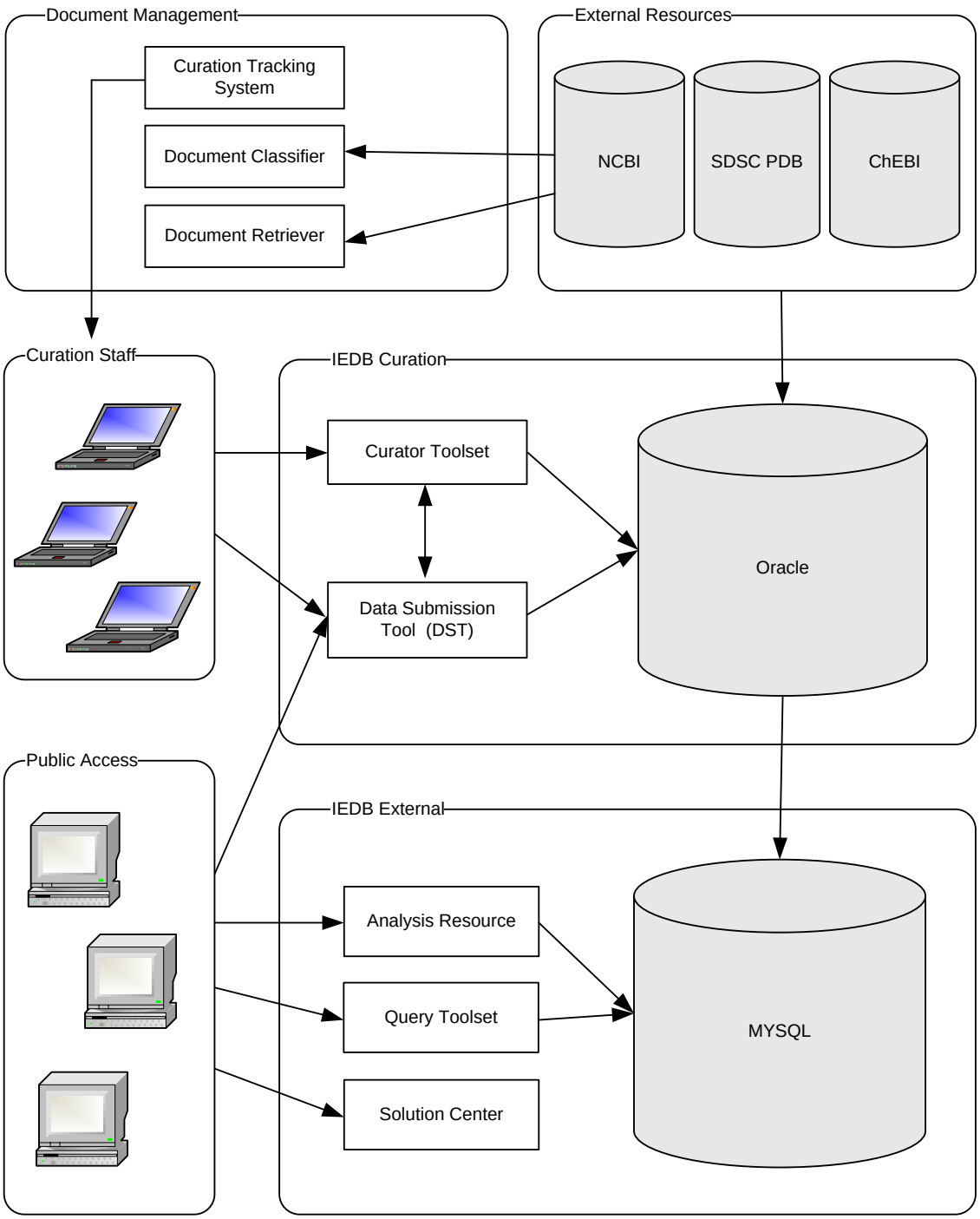

Figure 2-1. IEDB High Level Design Diagram

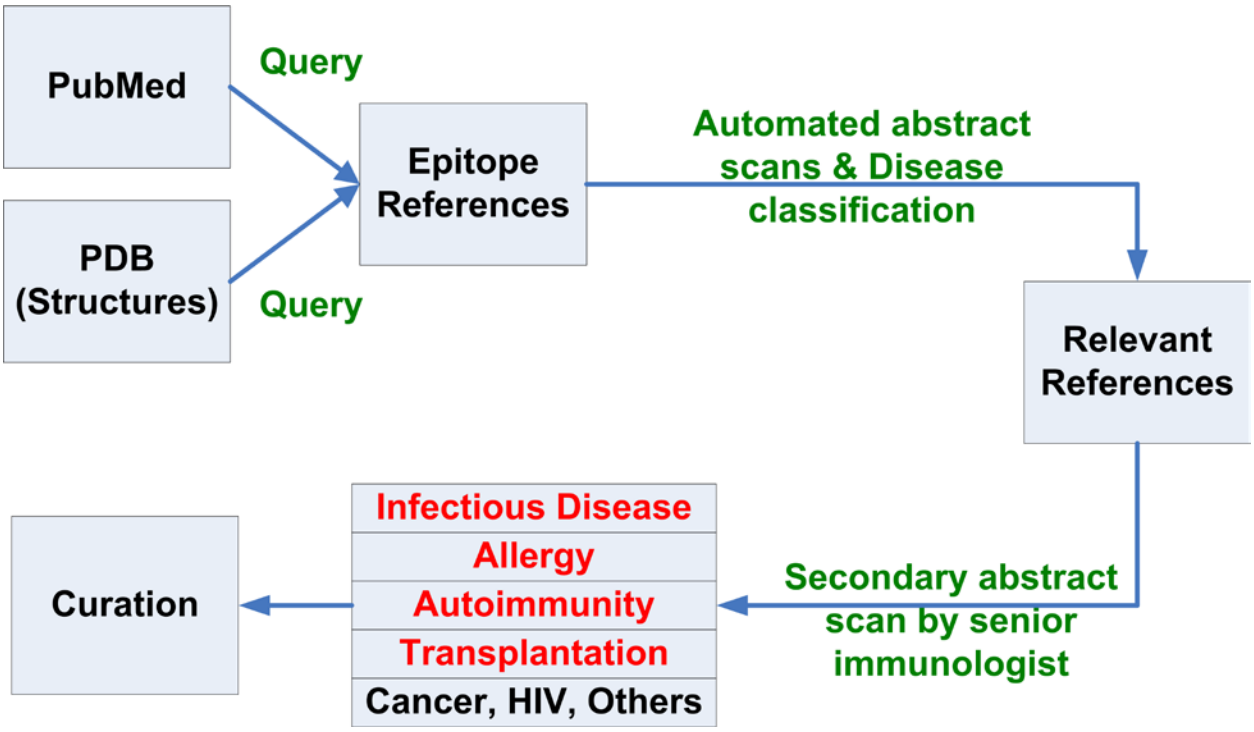

Figure 2-2. Overview Document Classifier Process

2.1.3 IEDB External Overview

The primary Website that is accessible to the public is the IEDB External website (http://www.iedb.org). The External application provides a query toolset which gives users the ability to query and download epitope data. This public site is linked to the IEDB Analysis Resource which contains tools for advanced analysis of epitope data including epitope prediction tools. The External site also includes the Solutions Center which provides support mechanisms to tutor, assist, and communicate with the user community. It also provides a portal for users to submit help requests and feedback. Section 3.7 contains the External data model information.

2.1.4 IEDB Analysis Resource Overview

The Analysis Resource is a collection of T cell and antibody epitope prediction tools and analysis tools. It resides on its own server, as described in Section 2.2. The IEDB website has help information, examples, and references for all the tools in the Analysis Resource. All the tools have been developed under the IEDB contract except for the tools developed by DTU. The DTU prediction tools were developed under separate funding and were delivered as executables for integration into the IEDB with the current IEDB contract funding. Source code for all tools except those developed by DTU and executables for the DTU tools will be made available to the follow-on entity during the transition period. Section 3.8 contains the Analysis Resource data model information.

2.1.4.1 T Cell Epitope Prediction Tools

The T cell epitope prediction tools fall into two categories – those that predict IC50 values for peptides binding to specific MHC class I and II molecules and those that predict epitope candidates based upon the processing of peptides in the cell.

For the MHC class I binding predictions, nine methods have been implemented, as listed below with their aliases. The asterisk indicates tools developed by DTU:

- Consensus
- netMHCpan*
- Artificial neural network (ANN)
- Stabilized matrix method (SMM)
- SMM with a peptide:MHC binding energy covariance matrix (SMMPMBEC)
- Scoring matrices derived from combinatorial peptide libraries (Comblib_Sidney2008)
- PickPocket*
- netMHCcons*
- netMHCstabpan*

The MHC class II binding prediction offers the user six different methods:

- Consensus
- NetMHCIIpan*
- NN_align*
- Stabilized matrix method align (SMM_align)*
- Combinatorial library
- Sturniolo

There are two general categories of T cell epitope processing tools. The tool in the first category combines predictors of proteasomal processing, TAP transport, and MHC binding to produce an overall score for each peptide's intrinsic potential of being a T cell epitope. The user can select one of seven methods:

- netMHCpan*
- Artificial neural network (ANN)
- Stabilized matrix method (SMM)
- SMM with a peptide:MHC binding energy covariance matrix (SMMPMBEC)
- Scoring matrices derived from combinatorial peptide libraries (Comblib_Sidney2008)
- PickPocket*
- netMHCcons*

The second processing prediction category contains two neural network tools. NetChop is a predictor of proteasomal cleavage sites and NetCTL predicts T cell epitopes along a protein sequence. Both tools were developed by DTU.

In addition, the Analysis Resource hosts MHC-NP, a tool that predicts peptides that are naturally processed by MHC. This tool was developed by Sébastien Giguère Alexandre Drouin, Alexandre Lacoste, Mario Marchand, Jacques Corbeil and François Laviolette.

2.1.4.2 B Cell Epitope Prediction Tools

There are three categories of antibody epitope prediction tools in the IEDB. The first is a collection of six methods that predict continuous antibody epitopes. Five of them use amino acid scales and the sixth, called BepiPred, predicts the location of linear epitopes using a combination of a hidden Markov model and a propensity scale method. BepiPred was developed by DTU. The second category contains DiscoTope, a tool developed by DTU that incorporates solvent-accessible surface area calculations and contact distances into the prediction of antibody epitope potential along the length of a protein sequence. This method can be used to predict discontinuous epitopes. The third category includes ElliPro, which predicts epitopes based upon solvent accessibility and flexibility. A fourth tool is available in this section of the website and can be used for modeling antibody structures when a PDB structure is not available. Prediction of ImmunoGlobulin Structure (PIGS) was developed by P. Marcatili and A. Tramontano.

2.1.4.3 Analysis Tools

The Analysis Resource has three tools that allow the user to further analyze a known epitope sequence or group of sequences:

- Population Coverage - This tool calculates the fraction of individuals predicted to respond to a given set of epitopes with known MHC restrictions. This calculation is made on the basis of HLA genotypic frequencies assuming non-linkage disequilibrium between HLA loci.

- Epitope Conservancy Analysis - This tool calculates the degree of conservancy of an epitope within a given protein sequence set at different degrees of sequence identity. The degree of conservation is defined as the fraction of protein sequences containing the epitope at a given identity level.

- Epitope Cluster Analysis - This tool groups epitopes into clusters based on sequence identity. A cluster is defined as a group of sequences that have a sequence similarity greater than the minimum sequence identity threshold specified.

2.2 Hardware Architecture

The IEDB has three basic systems (curation, external, and tools) that reside in three different locations. The curation system that is used for curating scientific literature and processing data submissions from external researchers consists of three nearly identical hardware and software environments. The three systems host a development system, a test system, and a curation production system, which is the one actually used by the curators. Table 2-1 describes the hardware components for the IEDB.

Table 2-1. IEDB Hardware Components

Component Name	Description
Firewalls	Protects servers from unauthorized access. Firewalls exist at LJI and SDSC to support networks based in those locations.
Database Servers	Database servers support the database software. There are twelve database servers used to support the IEDB; two database servers running Oracle for Curation (production and development), five database servers running MySQL for External (two local production, two remote production, and one local development), and five database servers running Oracle XE for the Finders, a search feature (two local production, two remote production, and one development).
Application Servers	Servers that support the application software. There are five application servers: two curation servers (production and development) and three tools servers (local production, remote production, and development).
Web Servers	The servers supporting the web software. There are five web servers are used by External (local production, remote production, and development).
Virtual Machine Host Servers	Virtual machine host servers are used to support the External software (production and development).
SAN Storage	All production and development machines are hosted in a local Storage Area Network at their respective geographical locations.

All servers are virtual machines hosted in a VMWare ESX vSphere environment and utilize a redundant SAN for storing the VMs. All local production virtual machines are replicated offsite to the San Diego Supercomputer Center (SDSC) excluding the Curation servers.

At LJI the Curation database and application servers are hosted on dedicated ESX hosts on HP BL460c G7 server blades. All other virtual machines are distributed amongst four ESX hosts on HP BL465c G7 blades. All of the storage for IEDB VMs are located on a dedicated SAN storage system.

The SDSC location currently utilizes SAN storage which is presented to a Supermicro Blade System running VMWare vSphere. There are four Supermicro BHDGT host compute blades for all VMs at SDSC.

The system that users see when they go to www.iedb.org is referred to as the external system. Physical machines are located at LJI and are replicated offsite to an offsite facility at SDSC in La Jolla, CA. The virtual servers in the production set are duplicated. One is used for staging the weekly update while the other actually serves as the production machine visible to the public. The staging and production environments are swapped weekly at the end of the update on the staging machine.

Figure 2-3 depicts the server configuration of the IEDB production environment. Section 2.3 lists and describes the software used to develop and maintain the IEDB. The hardware and software relationships are depicted in Section 2.4.

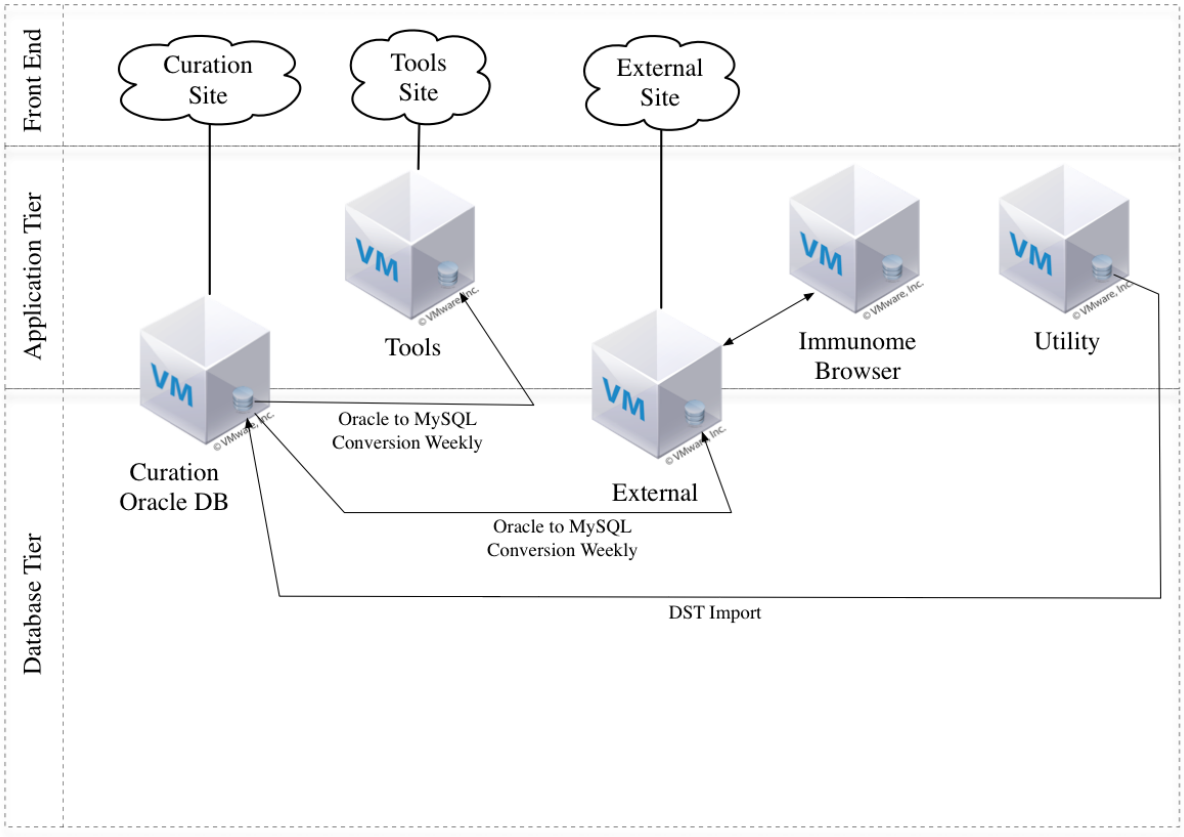

Figure 2-3. IEDB Production Environment

2.3 **Software Components**

Table 2-2 Describes the software components for the IEDB application.

Table 2-2. IEDB Software Components

Component Name	Description
AJAX	Asynchronous JavaScript and XML techniques for faster, more dynamic user interface. (External)
Ant	Apache Ant is a software tool for automating software build processes. (Curation and External)
Apache Axis	An open source, XML based web services framework. Used by internal curation system for retrieval of PubMed records. (Curation)
Apache Subversion (SVN)	SVN is an open source software versioning and revision control system. (Curation and External)
Apache Web Server	The Apache web server is responsible for handling HTTP requests, routing dynamic calls to either the Tomcat servlet container or PHP files. (External)
Display tag library	Open source jsp tag library specializing in tabular data presentation. (Curation)
Drools (JBoss Drools)	Drools is a business logic management tool, used to enforce data validation rules during the curation process. (Curation)
Dojo Javascript Toolkit	The Dojo Toolkit is an open source modular JavaScript library, used on forms, tree browsers and AJAX calls. (External)
Eclipse IDE	Eclipse is an open source Integrated Development Environment. (Curation and External)
Java Server Pages	High level abstraction of Java Servlets representing the 'View' layer of MVC architecture. (Curation and External)
MySQL / MariaDB Database	Open source RDBMS used by the public facing query system as well as analysis tools. MariaDB is a community-developed fork of MySQL (External)
Oracle Database 12c	RDBMS used by internal curation system (Curation)
PHP	Lightweight server-side web development language, used within Apache Web Server. (External)
Python	A general purpose high-level programming language used by the public facing query system as well as analysis tools. (External)
Spring Framework	An open source application framework for Java providing inversion of control, data access and transaction management. (Curation)
Struts	Apache (formerly Jakarta project) Struts is an open source MVC framework for developing J2EE applications. (Curation and External)
Tiles	Apache Tiles is a J2EE View framework, allowing JSP pages to be developed modularly. (Curation and External)
Tomcat	Apache Tomcat is an open source servlet container. It is used for logic heavy operations on the Public facing site. (External)
WebLogic Application Server	J2EE application server. Hosts web tier, application logic tier, and data access tier for internal curation system. (Curation)

Component Name	Description
Web Ontology Language (OWL)	Used to represent taxonomies for use by the finder applications (Curation and External)
XML	Extensible Markup Language is a set of rules for encoding documents in machine readable form. It is used in the internal curation systems import and export processes. (Curation)

2.4 Hardware and Software Relationships

Each software component is used in conjunction with one or more hardware components. The table below describes the hardware, operating systems and key software used with that hardware component.

Table 2-3. Hardware and Software Relationships

Hardware Component	Software Component
Application / Database Servers (Curation)	CentOS Enterprise Linux Server Oracle Enterprise RDBMS Sun Java WebLogic Server
Application Servers (Tools)	CentOS Enterprise Linux Server Sun Java Apache Tomcat Python Apache HTTP Server MariaDB PHP
Web / Database Servers (External)	CentOS Enterprise Linux Server Sun Java Apache Tomcat Python Apache HTTP Server MariaDB PHP
Virtual Machine Host Servers	VMWare vSphere

3.0 Database Design and Data Models

3.1 Overview of the IEDB Database Architecture

The main repository for immune epitope data is a relational database utilizing Oracle 12c as the management system and hosted on a Linux virtual machine. jThis repository is where all curation data is stored and maintained. There are three instances which exist to support the development, test and production curation sites. The curation production instance supports the curation staff and allows them to create and update curation data.

The curation database is normalized and converted to a MYSQL database which is hosted on a Linux virtual machine. The External or public website uses this MYSQL database. The de-normalization of the main curation database allows for faster query performance on the external web site. A normalized version of the External MySQL database is also created and is available for download on the Database Export page of the External site. The physical data model for this database is represented in Figure 3-11.

3.2 Overview of IEDB Database Design

From a design standpoint, it is essential to first understand the domain and scope of the IEDB system. The IEDB data structure has been refined many times since its initial creation. The system will continue to change in response to other efforts including tool development, curation of different types of epitopes, Epitope Discovery Group data submission, and community feedback.

3.3 Overview IEDB Database Content

The scientific domain is critical to the design of the IEDB database. Within the IEDB data architecture there are three major concepts; reference, epitope, and assay. The top level object within our database is the reference. The origin of all data published in the database must be cited. This information is captured as a reference. A reference can be a publication or a submission and will contain one or more epitopes. Each epitope may have any number of assays and corresponding immunizations. Each assay results in one data point or measurement. For example, if four assays were performed to test the affinity of an epitope to a given MHC molecule with different assay techniques, there would be four assay records, one for each reading. Within the database each of these major concepts is represented by several tables. Figure 3-1 depicts a high level overview of the IEDB data components and the associated relationships.

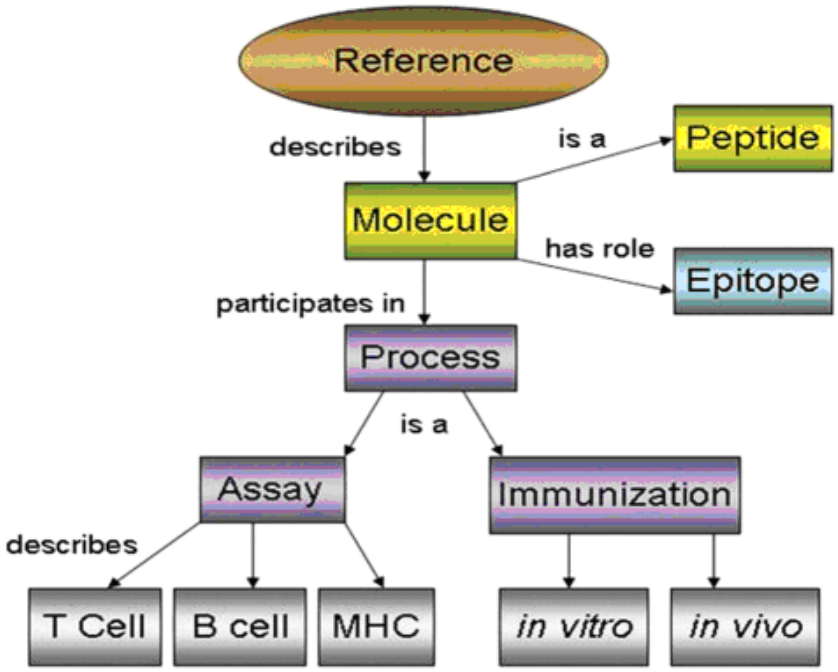

Figure 3-1. IEDB Database Content

Reference data are curated within an Oracle 12c database using the internal Curation IEDB site (http://curation.iedb.org/home.do). The Curation application is password protected and accessed only by Curators. A weekly build process creates a read-only, de-normalized MYSQL database from the curated references with have been promoted to production status. This MYSQL database is used by the External IEDB site (www.iedb.org) which is open to the general public. The sizes of the different databases are listed in Table 3-1 below.

Table 3-1. IEDB Database Size

Database	Size
CURATION ORACLE	143GB (20GB compressed)
EXTERNAL MYSQL	36GB (6GB compressed)
MYSQL IEDB_PUBLIC	3GB (350MB compressed)
IEDB_ANALYSIS MYSQL	100MB (50MB compressed)

3.4 IEDB Reference Data Flow

The diagram below depicts how references flow between the internal and external application databases. Using the internal Curation application, references are introduced from PubMed or via submissions into the NEWDB "staging area" schema inside the Curation Oracle database. When references are ready to be displayed on production, they are copied, or "promoted", to the NEWDB_PRODUCTION schema. A weekly build process recreates a set of de-normalized tables inside the NEWDB_PRODUCTION schema representing the production curated references. These tables are then transferred via the ETL utility PAN into the three MYSQL

databases which are recreated during the weekly build. If a production reference needs to be altered, the reference is copied from the NEWDB_PRODUCTION schema back into the NEWDB schema for re-curation. At the same time, the reference is copied to the NEWDB_COPY schema. This provides rollback capability so that any changes to the staging area copy of the reference can be restored to the previous version from production. Data are migrated between the three internal Oracle schemas via Oracle stored procedures. The IEDB_ANALYSIS MYSQL database is used by the Epitope Prediction and Analysis Tools website. The IEDB_QUERY_YYYYMMDD MYSQL database is used by the IEDB website. The IEDB_PUBLIC MYSQL database is used for MYSQL database exports on the IEDB website. The IEDB_PRIVATE are MySQL copies of the Oracle NEWDB and NEWDB_PRODUCTION tables that can be used for internal development.

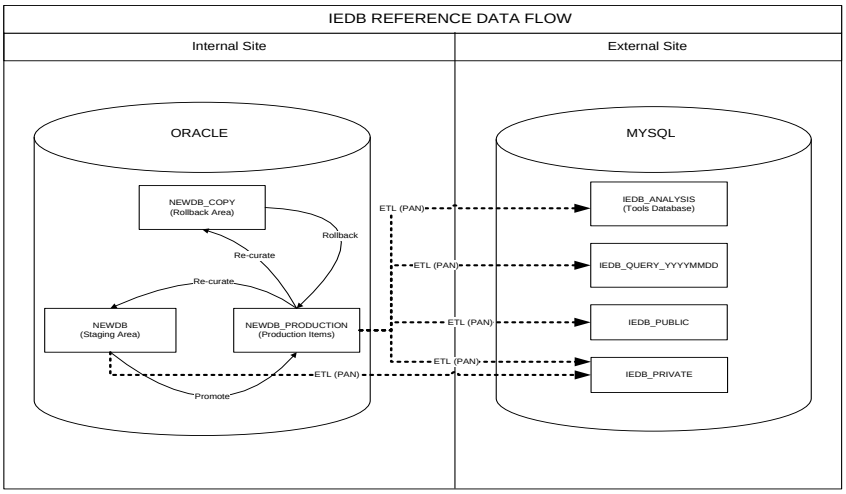

Figure 3-2. IEDB Reference Data Flow

3.5 IEDB Database Backup Procedures

The Oracle 12c database used to store curation data is backed up using a variety of procedures. A full set of "Hot" backups are performed daily for all the curated data in the NEWDB, NEWDB_COPY and NEWDB_PRODUCTION schemas using Oracle's export utility. Using a defined process, these backups can be used to restore the curation database if needed or used to refresh the test curation database to sync up test with production. Data files are also backed up as part of the SAN storage system backup plan.

3.6 IEDB Curation Physical Data Models

The following data models represent the key areas of the IEDB database. These tables represent how the data is physically stored in the IEDB.

3.6.1 IEDB Reference and Base Data Model.

Figure 3-3 represents the Reference table and other related base tables of the curation application. Note that for space considerations, assay tables and their relationships are not displayed in this diagram. See the following individual assay diagrams for their direct relationships to the base tables.

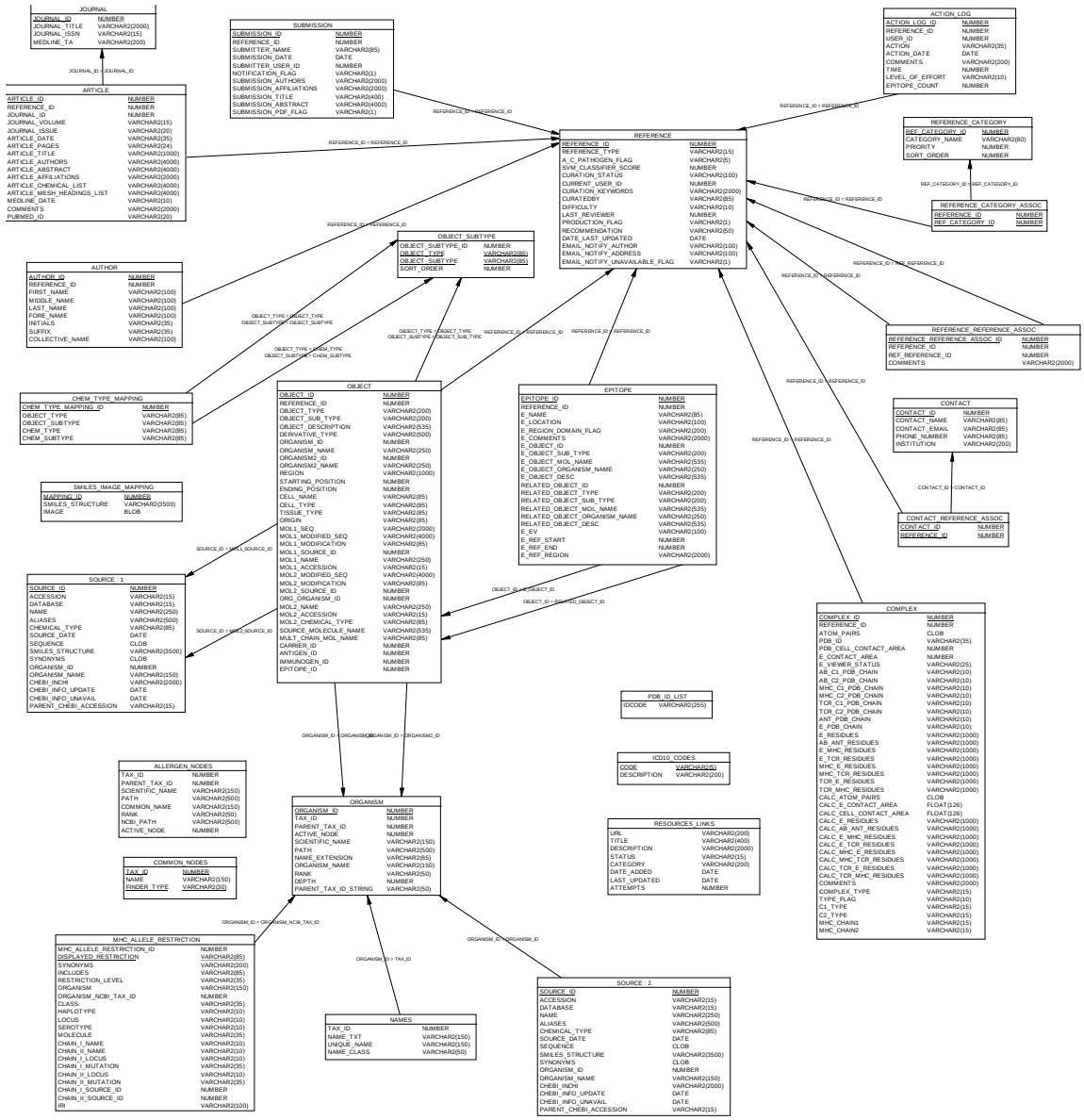

Figure 3-3. IEDB Reference and Base Physical Data Model

3.6.2 IEDB TCell Data Model

Figure 3-4 represents T cell assay data and its direct table relationships.

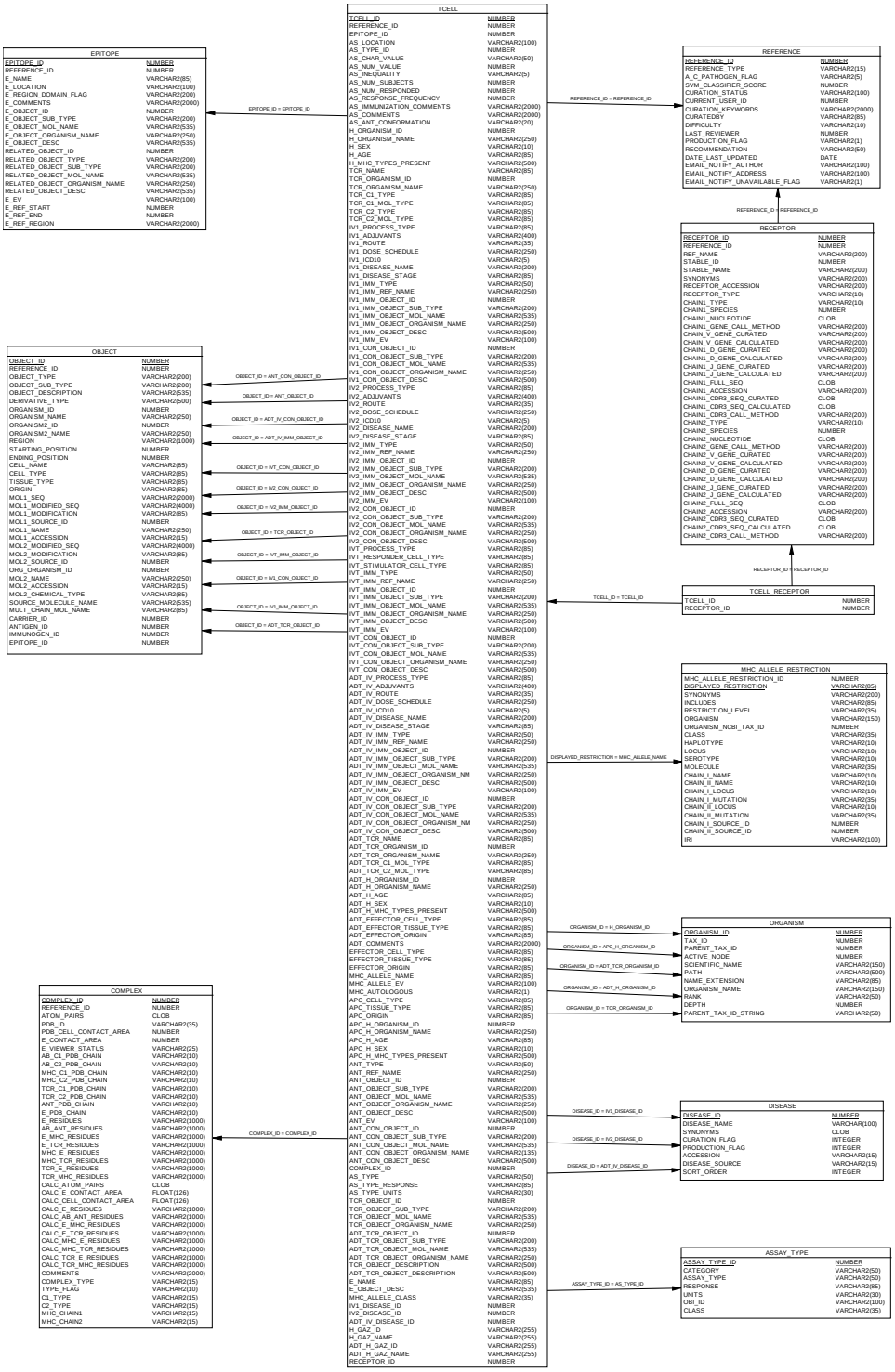

Figure 3-4. IEDB TCell Assay Physical Data Model

3.6.3 IEDB BCell Data Model

Figure 3-5 represents B cell assay data and its direct table relationships.

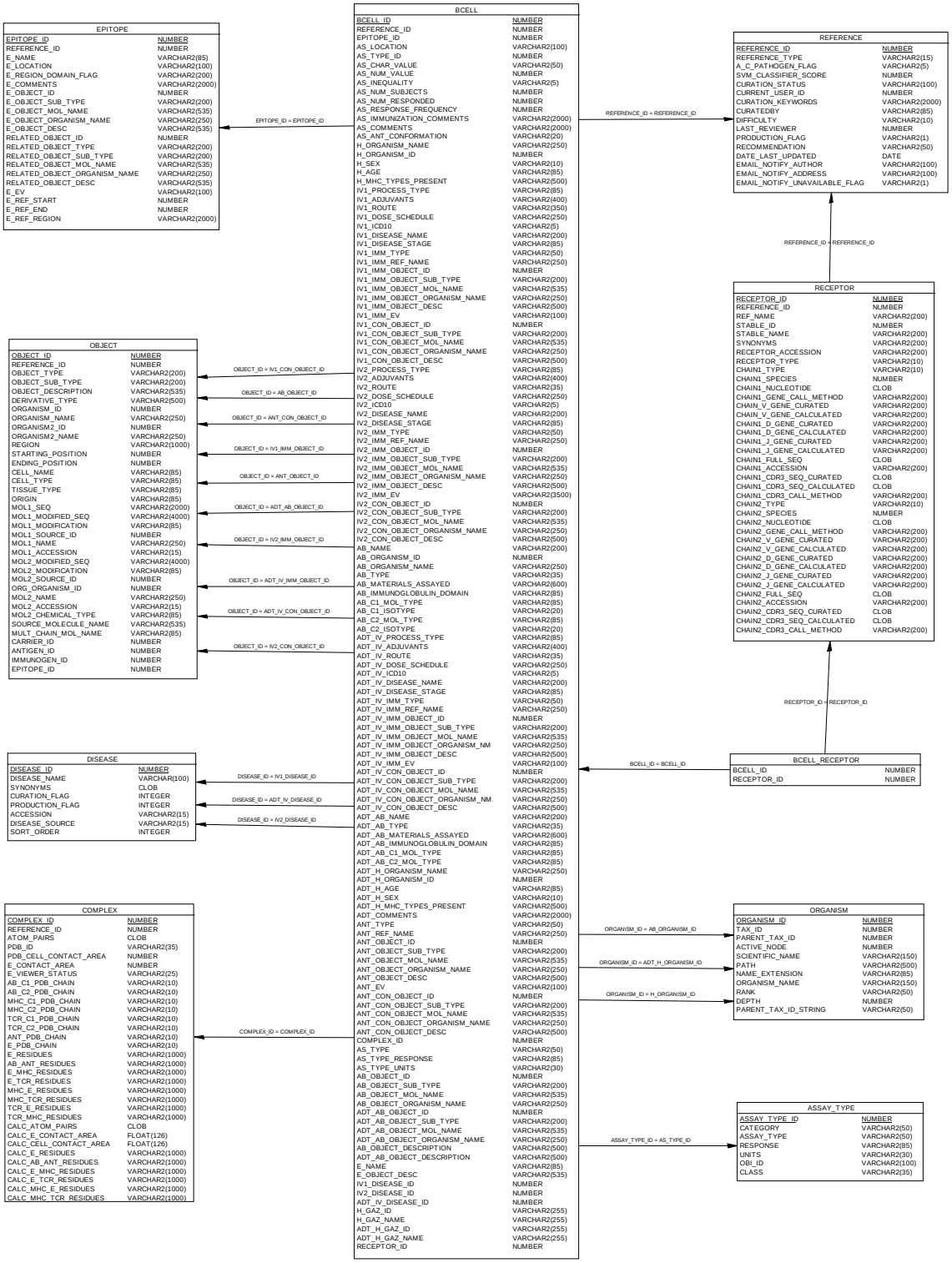

Figure 3-5. IEDB BCell Assay Physical Data Model

3.6.4 IEDB MHC Binding Data Model

Figure 3-6 represents MHC Binding assay data and its direct table relationships.

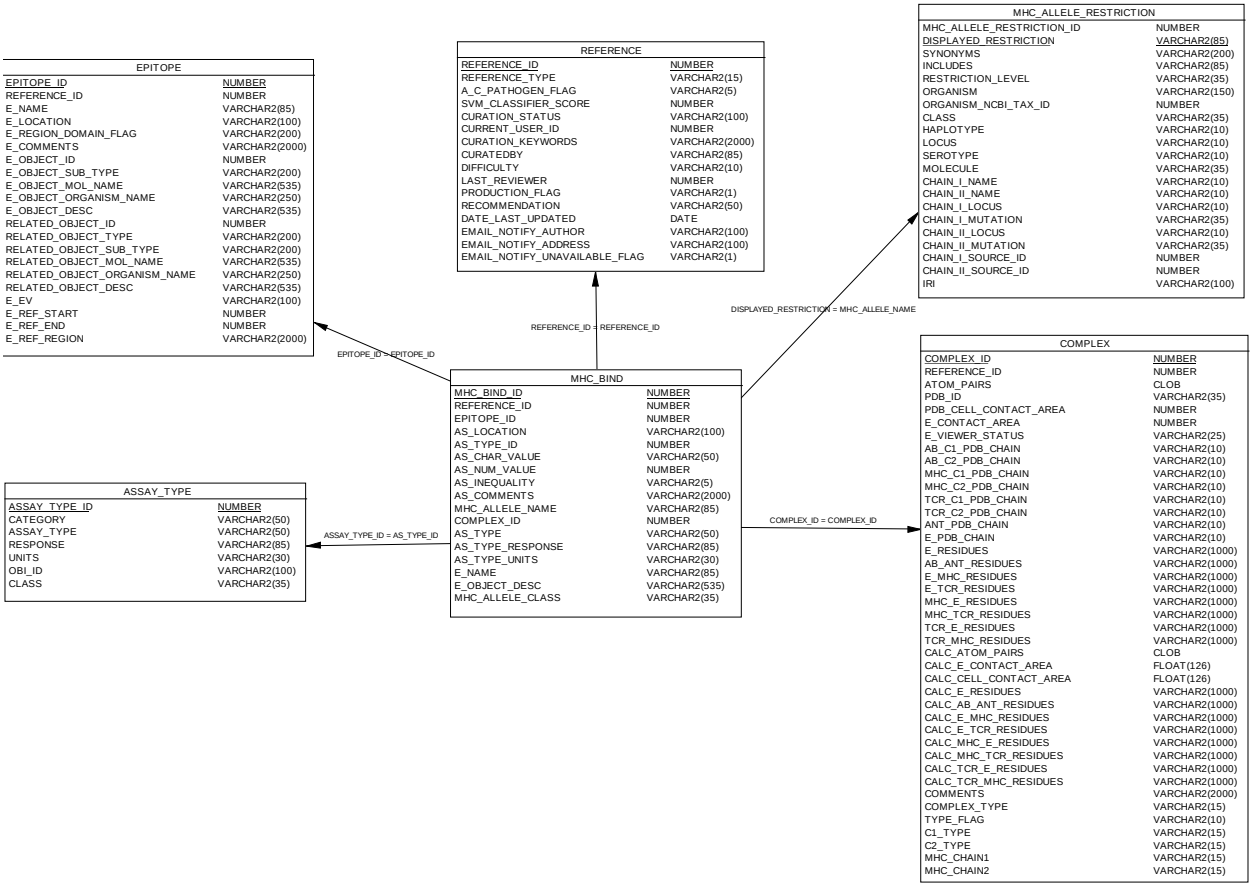

Figure 3-6. IEDB MHC Binding Assay Physical Data Model

3.6.5 IEDB MHC Ligand Elution Data Model

Figure 3-7 represents MHC Ligand Elution assay data and its direct table relationships.

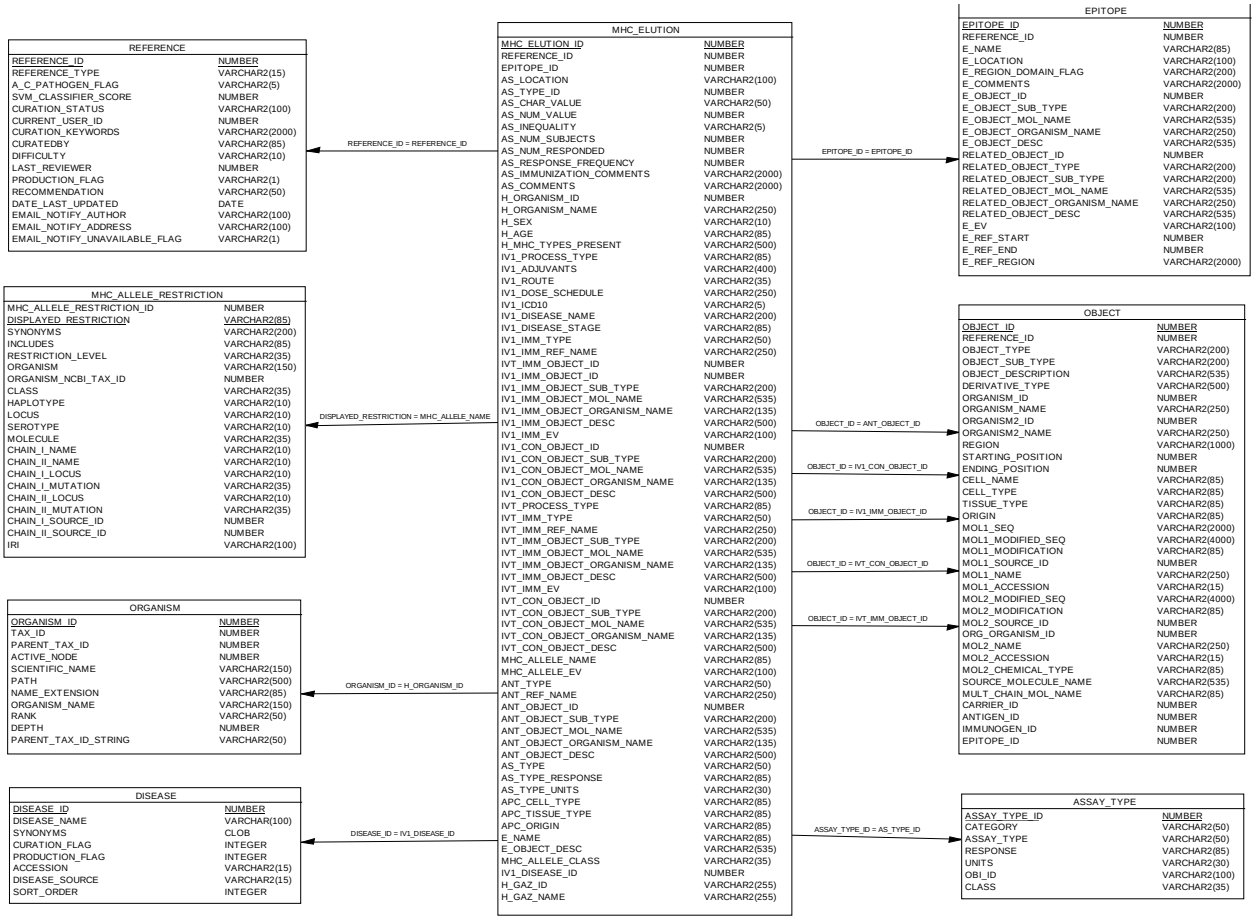

Figure 3-7. IEDB MHC Ligand Elution Assay Physical Data Model

3.6.6 IEDB ChEBI Support Data Model

Figure 3-8 represents ChEBI molecule data which is downloaded locally for use within the curation application.

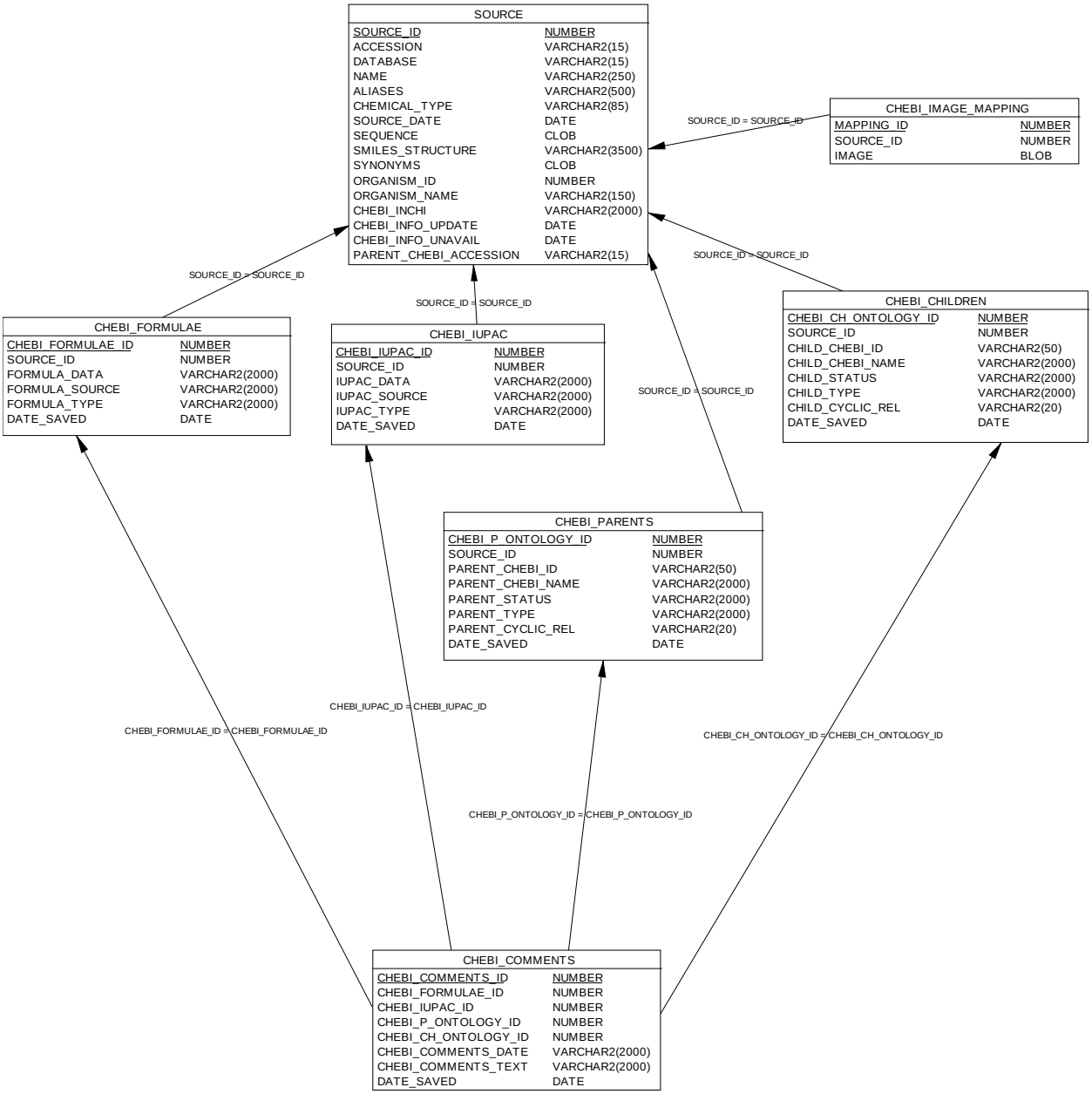

Figure 3-8. IEDB ChEBI Support Physical Data Model

3.6.7 IEDB User Data Model

Figure 3-9 represents the tables used for user authentication and authorization within the curation application.

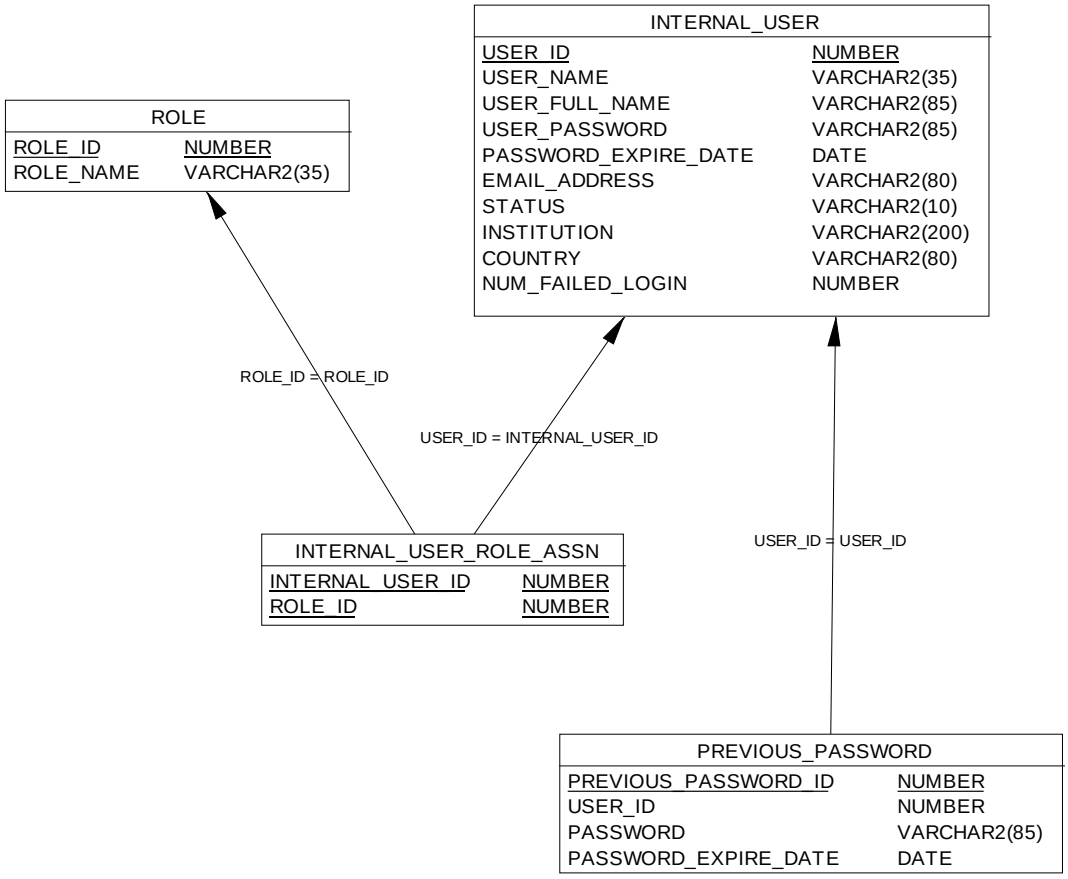

Figure 3-9. IEDB User Data Physical Data Model

3.6.8 IEDB Lookup Value Data Models

Figure 3-10 represents tables used to manage "list of values". These are used within list boxes on input forms within the curation application.

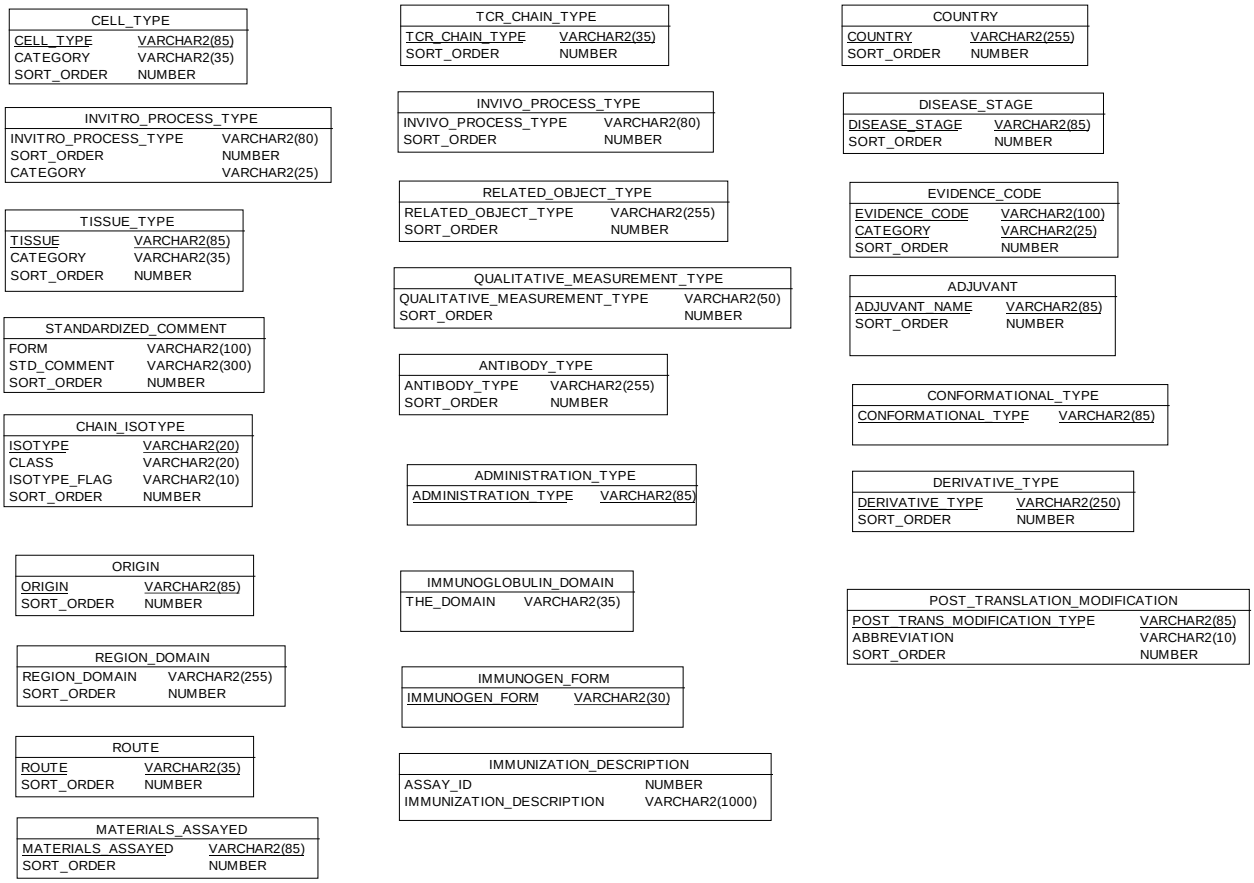

Figure 3-10. IEDB Lookup Value Physical Data Model

3.7 IEDB External Physical Data Model

Figure 3-11 represents the key areas of the IEDB External database. The following data model represents a normalized version of the External MySQL database.

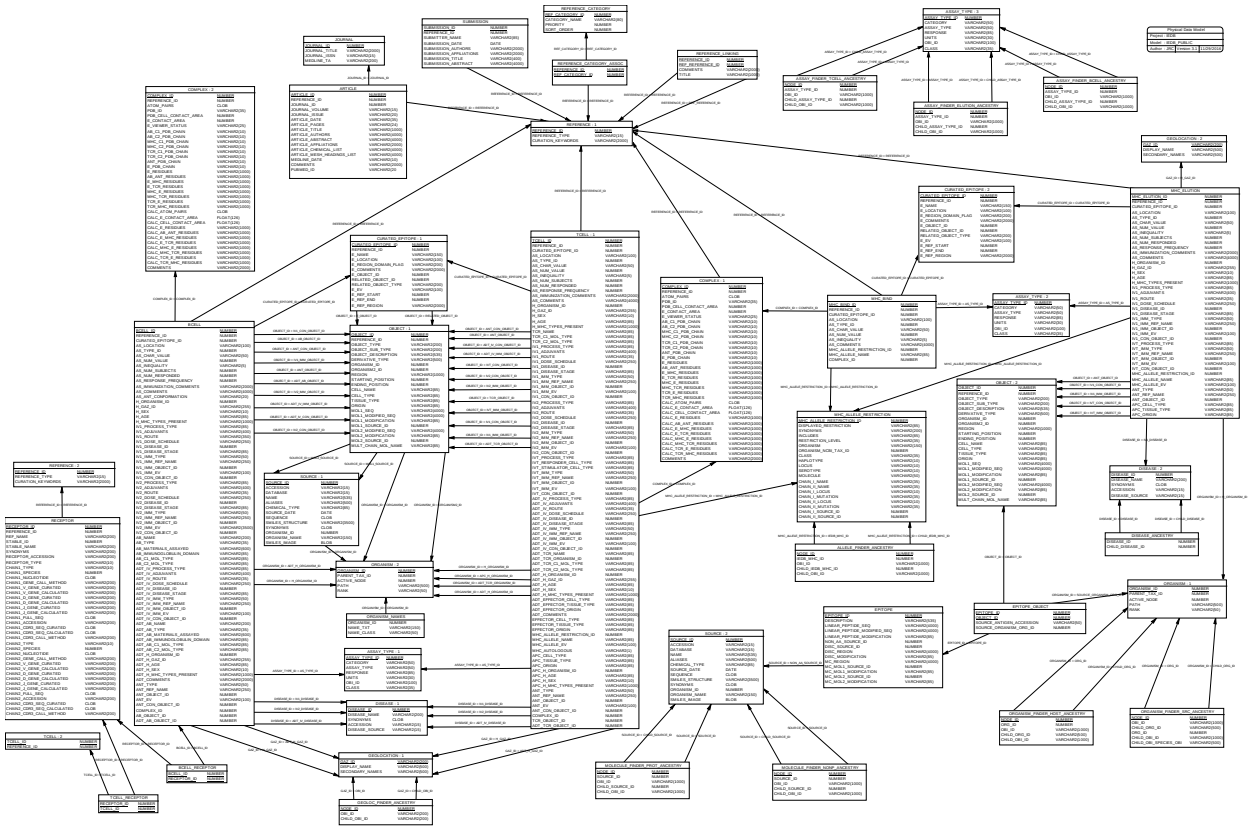

Figure 3-11. IEDB Public External Physical Data Model

3.8 IEDB Analysis Resource Data Model

Figure 3-12 represents the data which is generated for use within the Analysis Resource tools.

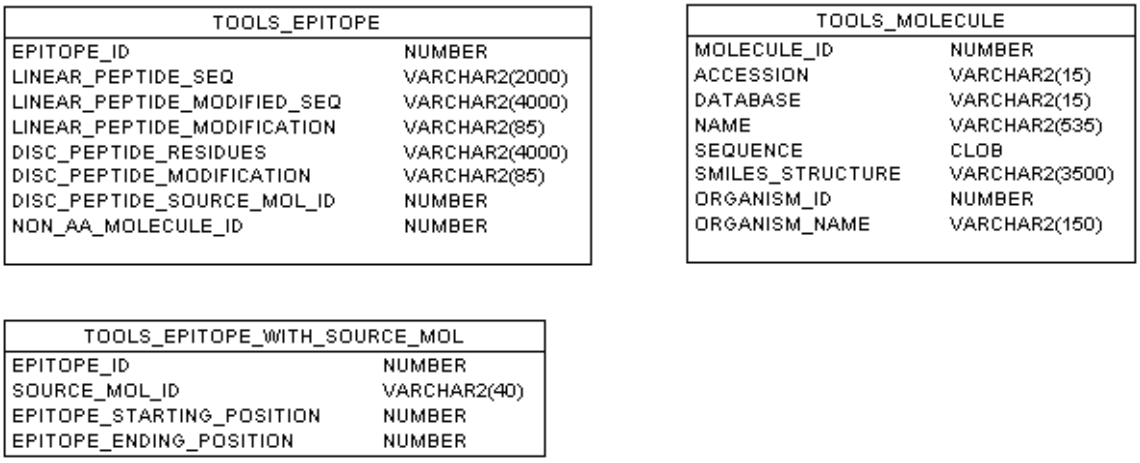

Figure 3-12. IEDB Analysis Resource Physical Data Model