User Guide for PDB-IHM Deposition and Data Harvesting System

A. PDB-IHM

PDB-IHM (formerly PDB-Dev) is a system for archiving structures of biological macromolecules determined using integrative modeling methods. Integrative modeling combines spatial restraints derived from a variety of biophysical techniques to determine the three dimensional structures of complex macromolecular assemblies. In addition, structures determined using integrative modeling can comprise of diverse spatio-temporal scales and conformational states, which pose unique challenges for data archiving.

Our goal is to create the infrastructure and software tools required to archive integrative models.

B. IHMCIF dictionary

We have extended the PDBx/mmCIF data standards used by the PDB to create the data representation for archiving integrative models. The extended data model is captured in the IHMCIF dictionary. Integrative structures archived in PDB-IHM comply with the IHMCIF dictionary.

IHMCIF extends the definitions in the PDBx/mmCIF dictionary in five significant aspects that address the requirements for archiving integrative models.

1. It allows for a flexible model representation with atomic and coarse-grained objects consisting of single and multi-residue spherical beads and three-dimensional Gaussian objects.

2. It supports constitutionally diverse structural assemblies and conformationally diverse ensembles, thereby providing representations for multi-state structural models and models related by time or other order.

3. It captures the spatial restraints derived from different kinds of biophysical techniques, such as CX-MS, SAS methods, EPR spectroscopy, DNA footprinting, mutagenesis, and others. Experimental restraints already captured in the PDBx/mmCIF dictionary and other related extensions are retained and reused where applicable. Several kinds of experimental data provide spatial restraints in the form of distances between atoms or residues (*eg*, distances from NMR NOE, FRET, and CX-MS experiments). To address the broad range of experimentally derived distance restraints, the IHMCIF dictionary includes a general representation of distance restraints between different kinds of features (*eg*, atoms, single and multiple residues,

contiguous residue ranges) and the corresponding uncertainties associated with these distance measurements. Representation of the spatial restraints addresses a key prerequisite for validating integrative models based on the experimental restraints.

4. It provides a generic representation for referencing related data from external resources *via* stable identifiers, such as accession codes or persistent digital object identifiers (DOIs). This is useful for referencing related data that either lives in an external repository (via stable accession codes) or does not yet have a primary repository (via standard DOIs).

5. It promotes reproducibility by incorporating simplified definitions for the modeling workflow and providing mechanisms to link modeling scripts and software program files.

The IHMCIF dictionary thus provides a comprehensive set of standardized definitions for representing multi-scale, multi-state, and ordered ensembles of complex macromolecular assemblies. The dictionary has been developed using diverse sets of examples and requirements gathered from the integrative modeling community.

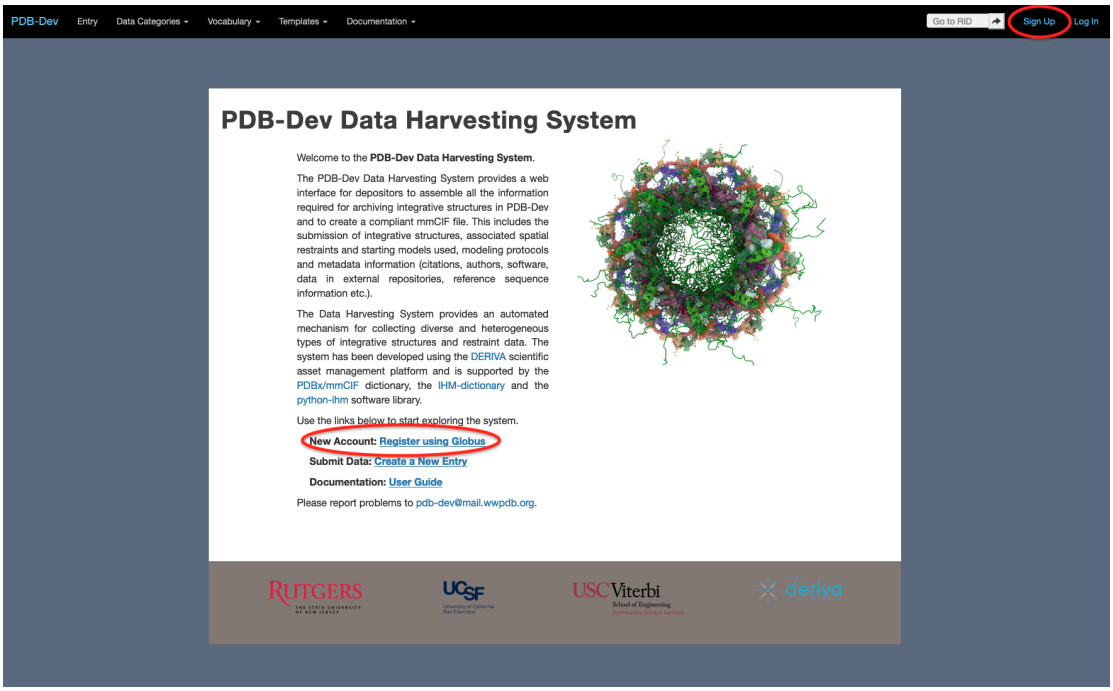
C. PDB-IHM Deposition and Data Harvesting System

The deposition and data harvesting system provides a web interface for depositors to assemble all the information required for archiving integrative structures in PDB-IHM and to create a compliant mmCIF file. This includes the submission of integrative structures, associated spatial restraints and starting models used, modeling protocols and metadata information (citations, authors, software, data in external repositories, reference sequence information etc.). The deposition and data harvesting system provides an automated mechanism for collecting diverse and heterogeneous types of integrative structures and restraint data.

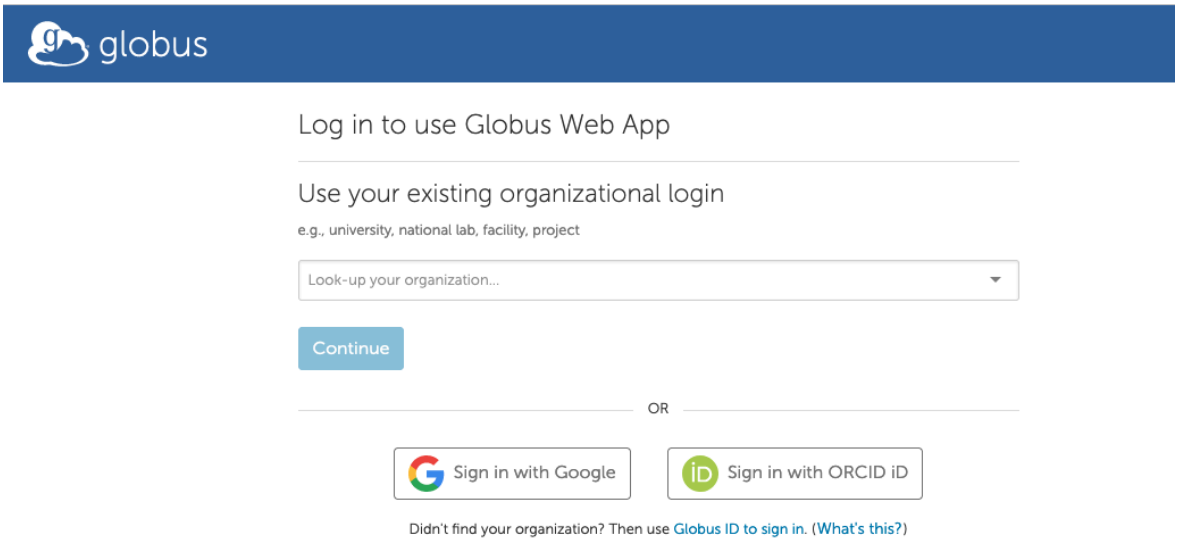
The system has been developed using the DERIVA scientific asset management platform and is based on the definitions in the IHMCIF dictionary and the parent PDBx/mmCIF dictionary. Most tables in the DERIVA catalog originate from a corresponding table in either the PDBx/mmCIF dictionary or the IHMCIF dictionary. Users can create a new entry in the catalog, upload the relevant data files and provide additional metadata information required for submitting an integrative structure to PDB-IHM. Once the user submits an entry, an mmCIF file compliant with the IHMCIF dictionary is generated. The following sections describe the details of using the deposition and data harvesting system.

D. Using the PDB-IHM Deposition and Data Harvesting System

1. Create an account (Globus)
    a. Go to https://data.pdb-ihm.org and click on the Sign Up link on the top right of the navigation bar or on the Register using Globus link on the webpage.

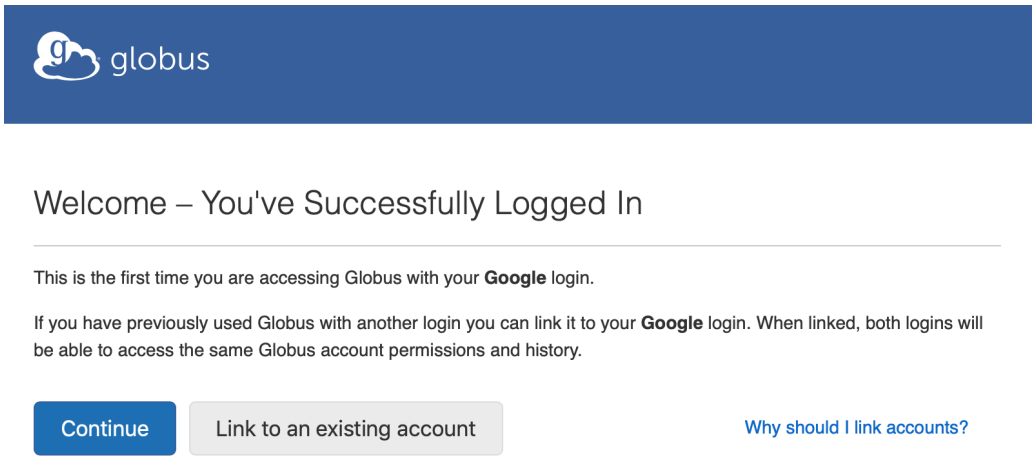


b. Choose how to login to Globus (using organizational account, google account or ORCID ID).

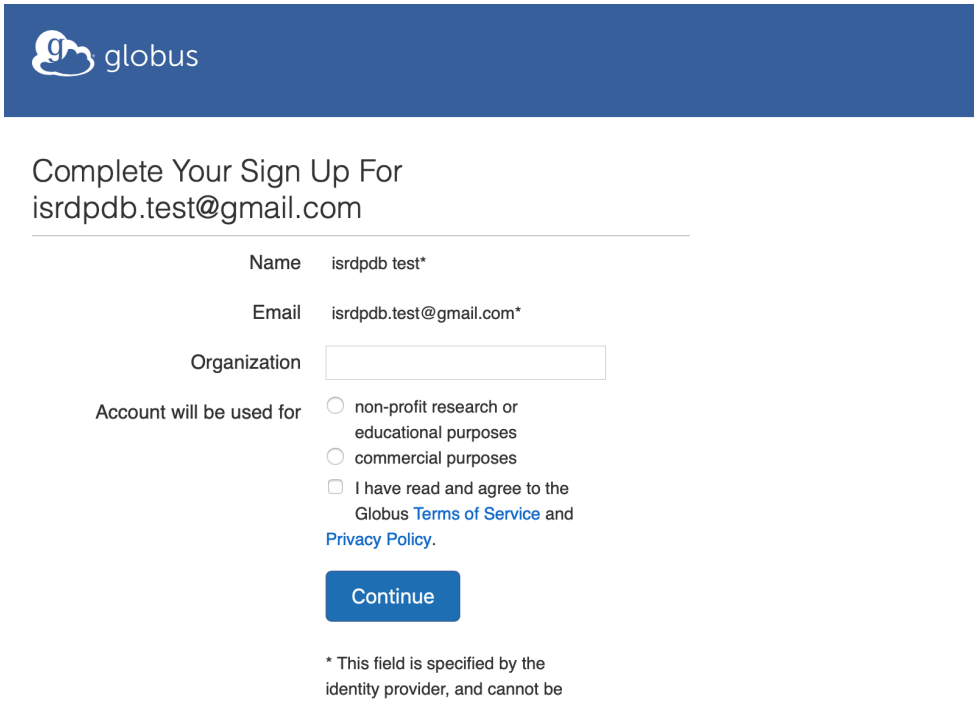


Note: If you see a page like the screenshot below and have an account you've used before with Globus, go ahead and click "Link to existing account". If not, click "No

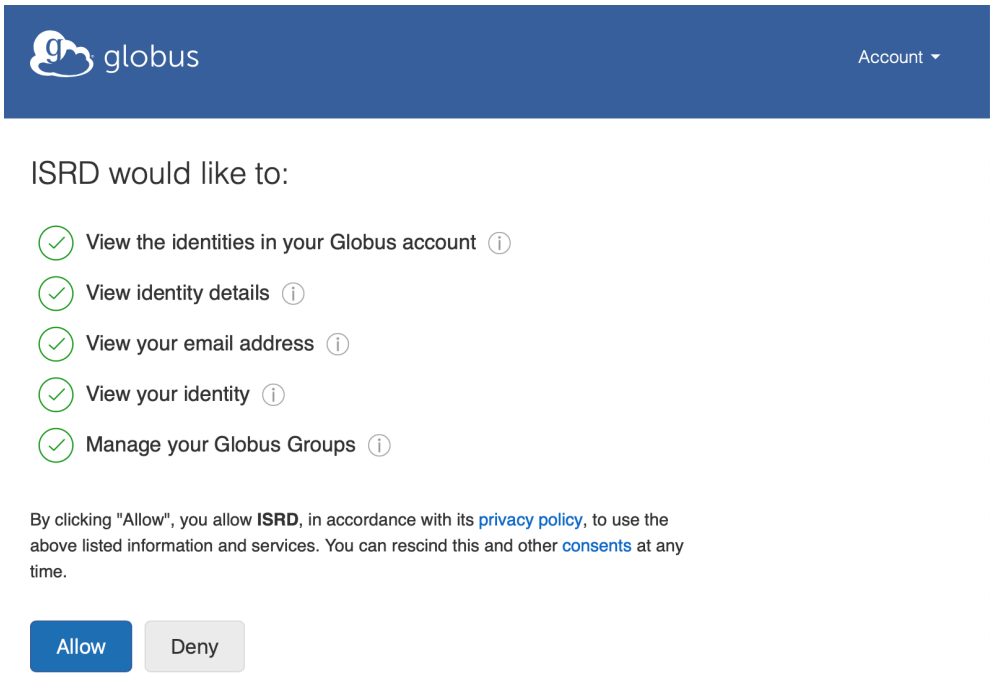
thanks, continue." (Note that if you have logged into Globus before and login with the same credentials, you'll skip directly to Step e.)



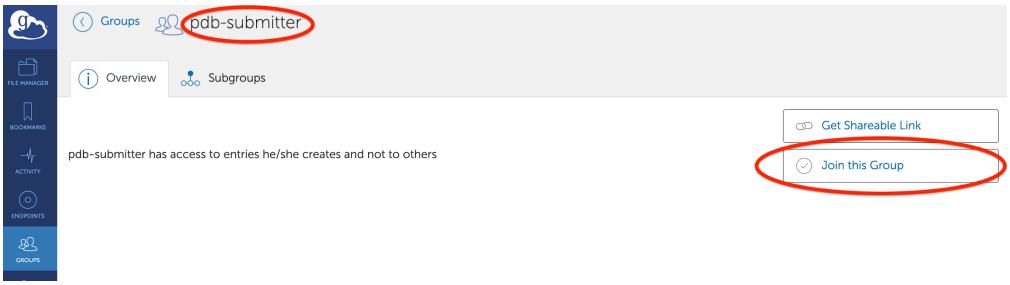
c. Complete a short signup form



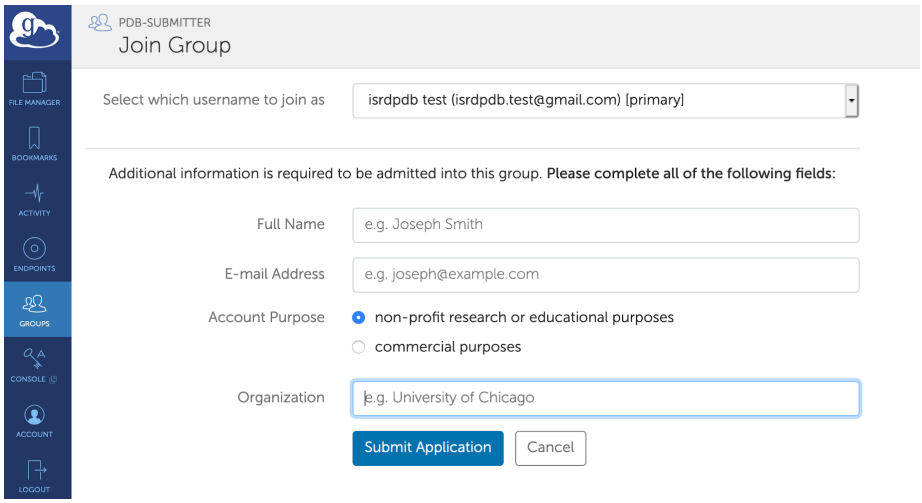
d. Provide consent to access your credentials



e. Once you reach the “pdb-submitter” globus group webpage, click on the “Join this Group” link to request to join the “pdb-submitter” group.



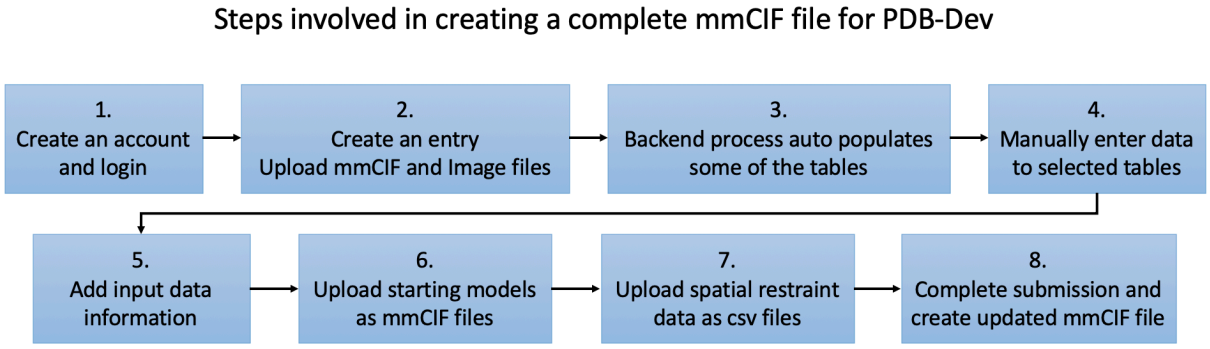
f. Fill in the form and submit the application to join. This will send a request to the administrators, who will review the credentials and approve the request.



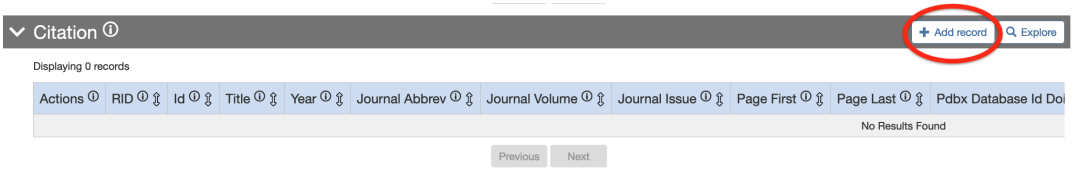
2. Login

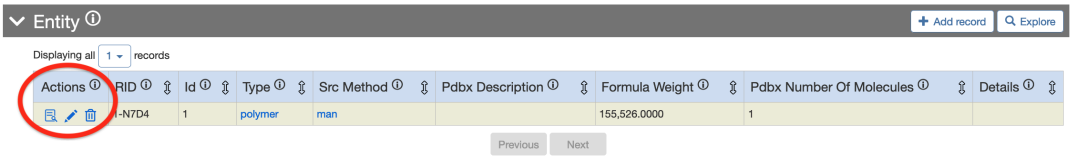
   Once the administrator has approved your join request, go to https://data.pdb-ihm.org and login using the credentials used to set up the globus account.

3. Workflow outline for data submission

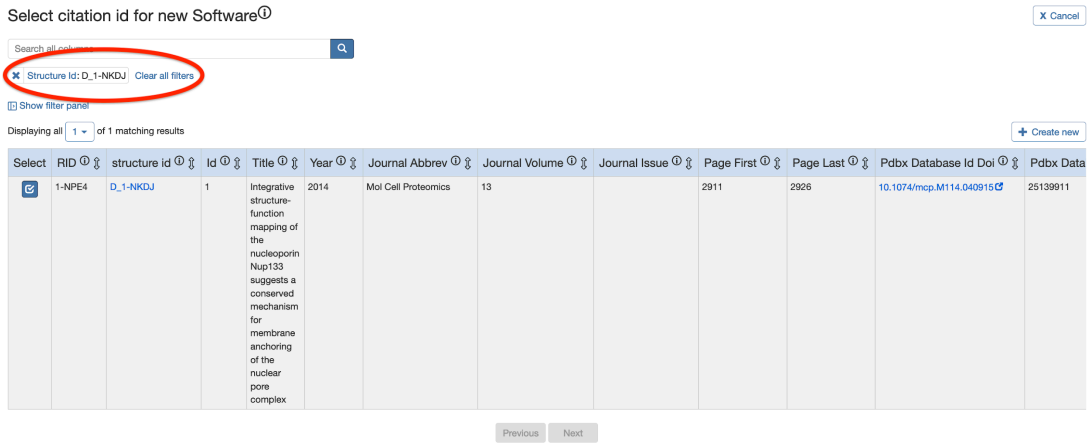


4. General information about using the system
   - An “id” or an “ordinal” column means an identifier for a row in a table and cannot accept duplicate values. The “id” can sometimes take unique text values, but sometimes requires unique integers. It is preferable to use positive integer values that increase systematically for such “ids” and “ordinals”. Work is under planning to autofill these integer “ids”.
   - To add a row to the table, click on the “Add record” link on the top right of the table. To view/edit/delete an already existing row in a table, click on the buttons under the “Actions” column on the left of the row.





- Do not clear “filters” on top left when entering form data to fill tables. These filters appear when referencing data from other tables. We are working on removing the option to clear filters.

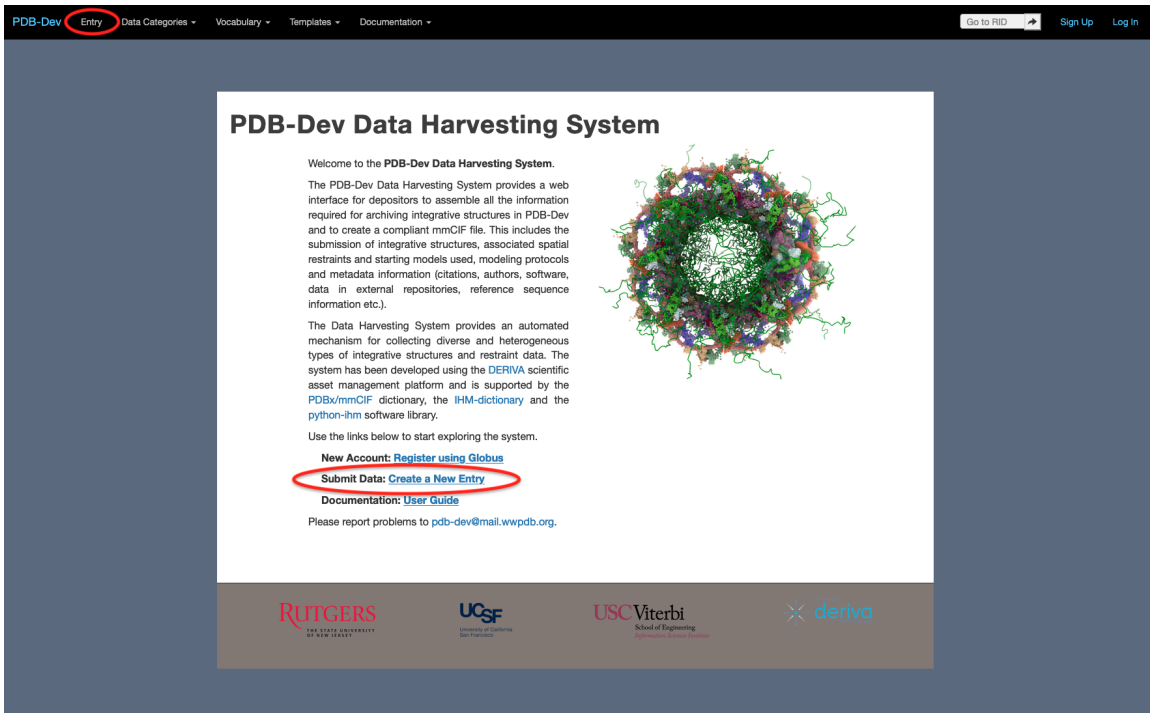


- There are several tables that are linked to each other through foreign keys (or parent-child relationships). This means that there is an order to fill up tables. The parent tables have to be populated before the child tables. The rest of this documentation takes this ordering into account and the instructions provided below make sure that the parent tables are filled first. For example, the “ihm_dataset_list” table that lists all the input data used in the modeling should be filled before adding details regarding the starting structural models used because the “ihm_starting_model_details” table has a foreign key pointing to the “ihm_dataset_list” table. This foreign key indicates which input data item in the “ihm_dataset_list” table corresponds to the starting model described in the “ihm_starting_model_details” table.

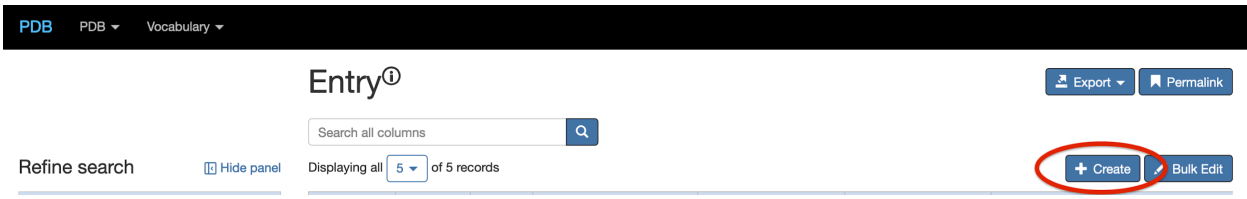
- Access controls have been implemented so that users can only view entries that they have created and own.
- For general information regarding using the “Chaise” data browser, please refer to the chaise documentation.

Steps for submitting models and restraints

5. Submission Step 1: Create new entry and upload mmCIF and image files
   a. After logging into https://data.pdb-ihm.org, click on the link to Create a New Entry on the website or on the Entry link on the navigation bar.

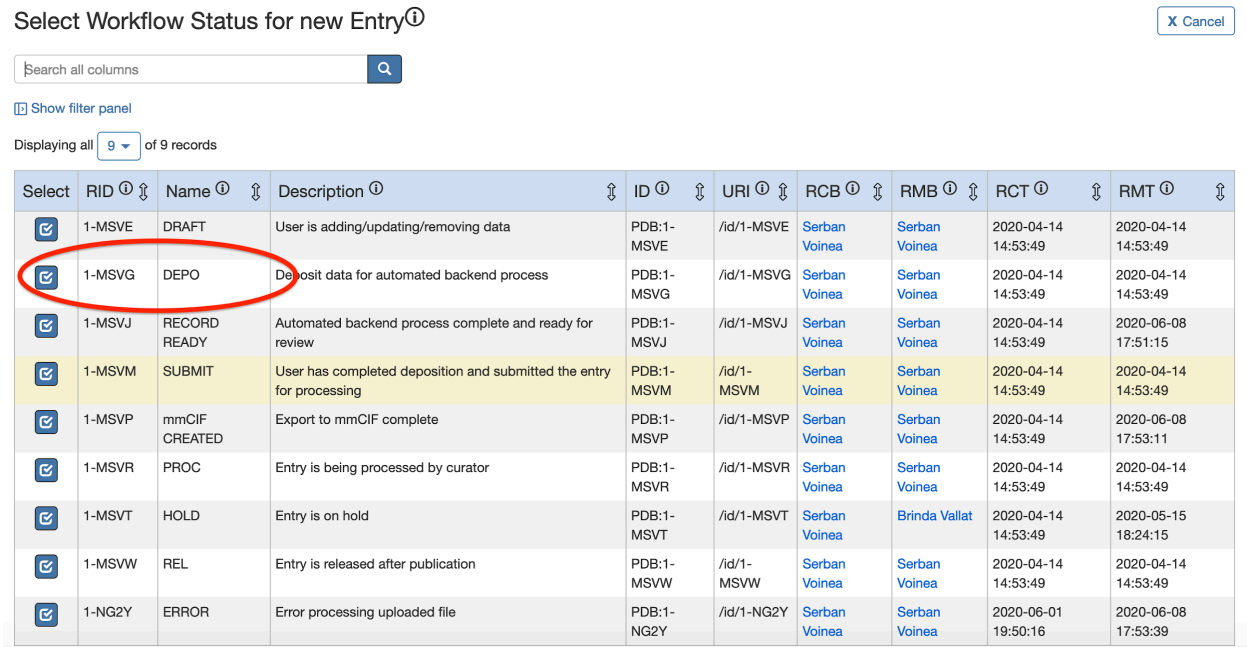


   b. The Entry page shows all entries created/owned by the user. Click on the “+Create” button on the left, near the top.

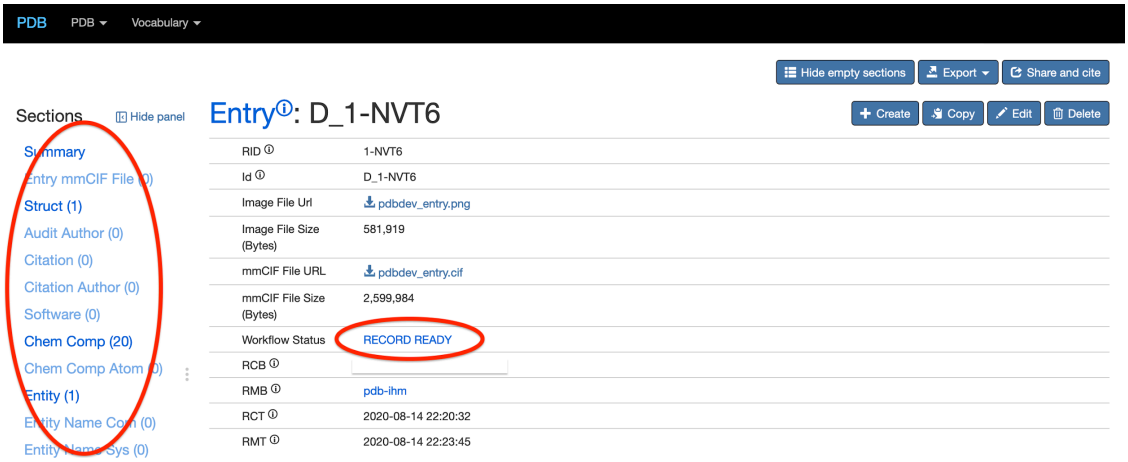


   c. On the new page, upload mmCIF file and Image file.

d. Choose the “Workflow Status” option as “Depo” and click “Save” on the form. This initiates the backend process to check the uploaded mmCIF file.



e. The backend process can take anywhere between a few minutes to a few hours depending upon the size of the mmCIF file uploaded. Once the backend process is complete, the “Workflow Status” for the entry is updated to “Record Ready”. Please wait till the “Workflow Status” is updated from “DEPO” to “Record Ready” before proceeding.



f. On the sidebar in the above figure (under “Sections”), are the list of mmCIF tables relevant for an IHM deposition. The backend process uses the uploaded mmCIF file to populate some of the tables in the Deriva catalog. The table names that begin with “ihm” are specific to integrative modeling.
   i. System Generated mmCIF file (not editable): Table populated with system generated final mmCIF file
   ii. struct (editable): High level table with information about the structure
   iii. chem_comp (not editable): List of chemical components (ligands and monomeric components)
   iv. entity (editable): List of polymeric and non-polymeric entities in the structure
   v. entity_poly (not-editable): Polymeric entities
   vi. entity_poly_seq (not-editable): Sequences of polymeric entities
   vii. atom_type (not editable): Types of atoms in the structure
   viii. struct_asym (not-editable): Instances of entities, both polymeric and non-polymeric
   ix. ihm_entity_poly_segment (editable): Segments of polymeric entities; specifies sequence ranges for use in other tables.
   x. ihm_struct_assembly (editable): List of structural assemblies
   xi. ihm_struct_assembly_details (editable): Details of structural assemblies; specifies polymeric (including segments) and non-polymeric components
   xii. ihm_model_representation (editable): List of model representations
   xiii. ihm_model_representation_details (editable): Details of model representation; addresses representations of multi-scale models with atomic and coarse grained representations

xiv. ihm_modeling_protocol (editable): List of modeling protocols
xv. ihm_model_list (not editable): List of models
xvi. ihm_model_group (editable): Groups of models (perhaps belonging to a state, cluster, etc.)
xvii. ihm_model_group_link (editable): Assignment of models that belong to a group

Some of these tables can be manually edited to add / modify data in subsequent steps, whereas some tables cannot be manually edited. For modifying tables that cannot be manually edited, the user needs to revisit step 1 and upload a new mmCIF file that re-initiates the step 1 automated backend process.

g. If uploading mmCIF and image files leads to an “ERROR” in the “Workflow Status”, users should NOT proceed to the next step until the errors are fixed. Users can try to fix the error based on the information available under “Record Status Detail”. Files can be re-uploaded and the workflow can be re-triggered if none of the tables on the sidebar are populated. If the tables are populated, a new entry can be created with updated files and the existing entry can be deleted. Users can contact the PDB-IHM team for any assistance.

6. Submission Step 2: Add/modify data manually

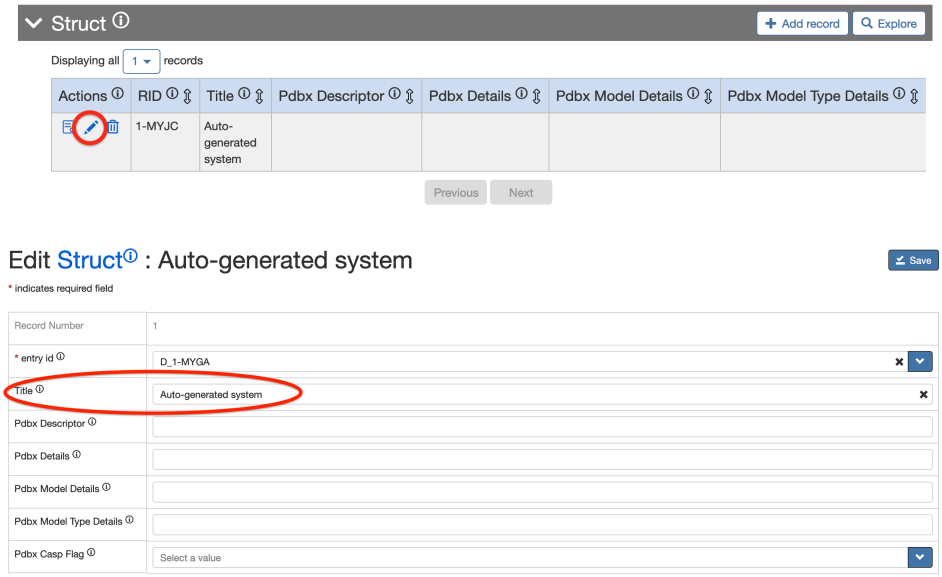
Once the “Workflow Status” in the “Entry” table is updated to “Record Ready” (see figure in section 5e above), the user can move on to Step 2, which is manual addition / modification of data.

There are several tables in the catalog where data needs to be added manually. Some of these are partially populated in step 1 during the automated processing. These partially populated tables can be further edited during this step to add / modify data. Information regarding what each table / column corresponds to can be obtained from the catalog using the “i” icon next to the table and column names. To add a row in a table, click on the “Add record” button on the top right of a table.

The topics and corresponding tables that allow for manual entry are listed below. It is preferable if data is entered into the tables in the order given below.

i. Entry and structure
    1. entry: Main entry table populated in step 1
       Note: System Generated mmCIF File is a table populated with mmCIF file generated by the system and is not editable.
    2. struct: Information regarding the submitted structure
        a. The backend process in step 1 provides a generic title for the structure (Auto-generated system). Click on the edit button next to that row in the “struct” table to update the

“title” for the structure.



ii. Authors of the structure
    1. audit author: List all authors of the submitted structure
iii. Citation information
    1. citation: List all citations related to the entry
    2. citation author: List all authors corresponding to the citations
iv. Software information
    1. software: List all software used in the modeling
v. Molecular entities in the structure deposited
    1. entity: Molecular entities in the structure. This table is partially populated in step 1. However, the table is editable to allow depositors to add descriptions and other relevant information regarding the molecular entities. Provide a name or description for the entity in the column “Pdbx Description”.
    2. entity_name_com: Common names associated with the entities
    3. entity_name_sys: Systematic names associated with the entities
    4. entity_src_gen: Details of the source from which genetically manipulated entities are obtained
    5. struct_ref: Details of reference sequence information from UNIPROT and GENBANK
    6. struct_ref_seq: Details of alignment information with the reference sequences from UNIPROT and GENANK
    7. struct_ref_seq_dif: Details of point differences in the alignment with the reference sequences from UNIPROT and GENBANK

8. pdbx_entity_nonpoly: Details regarding the non-polymeric entities in the structure

Note: Details regarding polymeric entities and chemical components are handled in step 1.

vi. Input datasets
   1. ihm_dataset_list: List all types of experimental data used in the integrative modeling study. Include all starting structural models as well as templates of comparative models. For example, chemical crosslinking (CX-MS), three dimensional electron microscopy (3DEM), experimental structures from the PDB
   2. ihm_dataset_group: List groups of datasets used in different steps in the modeling protocol, post processing, validation etc.
   3. ihm_dataset_group_link: Assign datasets to groups
   4. ihm_data_transformation: Details of rotation matrix and translation vector that can be applied to transform data
   5. ihm_related_datasets: Provide information about related datasets where one is derived from the other (e.g., 3DEM maps archived in EMDB derived from EM raw micrographs archived in EMPIAR).

vii. Data residing in external repositories
   1. ihm_dataset_related_db_reference: Provide accession codes for data archived in other repositories such as PDB, BMRB, EMDB, EMPIAR, SASBDB, PRIDE, MODEL ARCHIVE, MASSIVE etc.)

viii. Referencing data using DOI (data not in repositories)
   1. ihm_external_reference_info: Data that are not archived in other repositories can be referenced via a Digital Object Identifier (DOI). For instance data can be hosted in resources such as Zenodo or Figshare and the DOI is provided in this table.
   2. ihm_external_files: Resources like zenodo allow for users to store multiple files related to a research project under a single DOI. Provide names and paths for files archived within a single DOI.
   3. ihm_dataset_external_reference: Link the datasets used in the integrative modeling study (from the ihm_dataset_list table) to the files from the external resource such as Zenodo.

ix. Polymeric entity segments used in model representations, structure assemblies and starting models
   1. ihm_entity_poly_segment (partially populated in step 1): Define all polymeric segments (include sequence ranges) that are used in model representations (different segments that may have atomic and coarse grained representations), starting models (different segments that use different starting models - experimental, comparative or ab initio models), and structural assemblies corresponding to the models submitted.

x. Structure assembly

1. ihm_struct_assembly (partially populated in step 1 based on the models submitted): High level table to list all structural assemblies corresponding to the submitted models. If the submitted models are all constitutionally different, include an assembly corresponding to each model.
2. ihm_struct_assembly_details (partially populated in step 1 based on the models submitted): Compositional details corresponding to each assembly.
3. ihm_struct_assembly_class (used to represent hierarchical assemblies): List the different structural assembly classes that may belong to structural or functional hierarchies.
4. ihm_struct_assembly_class_link (used to represent hierarchical assemblies): Assign the assemblies that belong to the structural classes.

xi. Starting models used

1. ihm_starting_model_details:
    a. Add details of starting models used in the integrative modeling study. These starting models can be experimental models archived in the PDB, comparative models based on templates in the PDB, integrative models archived in PDB-IHM or ab initio models.
    b. Coordinates of starting models can be uploaded as mmCIF files. The chain identifiers and residue numbers in these mmCIF files should match the data entered into the table.
2. ihm_starting_comparative_models: Provide details regarding the comparative models (including template information) used as starting models in the integrative modeling study.
3. ihm_starting_computational_models: For purely computational models used as starting models, provide information regarding software used in generating these models.
4. ihm_starting_model_seq_dif: For experimental models used as starting models, provide details of point mutations between the starting models and the experimental models archived in the PDB (e.g., MET to MSE).

xii. Model representation

Model representation tables address the representation of multi-scale models that might consist of both atomic and coarse-grained representations.

1. ihm_model_representation (partially populated in step 1): High level table to list all model representations used in the integrative modeling study
2. ihm_model_representation_details (partially populated in step 1): Provide details of each model representation listed above

xiii. Modeling protocol details

1. ihm_modeling_protocol (partially populated in step 1): High level table to list all modeling protocols used in the integrative modeling study
2. ihm_modeling_protocol_details: Provide details pertaining to the protocols listed above including different steps involved in the modeling protocol
3. ihm_modeling_post_process: List all post processing steps carried out after the modeling to analyze the output models (e.g., filtering, clustering, etc.)

xiv. Models submitted

1. ihm_model_list: Populated in step 1 based on the models submitted; not editable
2. ihm_model_group (partially populated in step 1): Allows for structural models to be grouped based on clusters, states, time order or other order. List all groups or collections of models submitted.
3. ihm_model_group_link (partially populated in step 1): Assign models to model groups
4. ihm_model_representative: Identify the representative model in each model group
5. ihm_residues_not_modeled: Identify residues that are defined in the structural assembly, but are missing in the 3D model
6. ihm_multi_state_modeling: Describe the different states involved in a multi-state modeling study, if applicable
7. ihm_multi_state_model_group_link: Assign model groups to the different states defined above
8. ihm_ordered_ensemble: Provide details of ensembles related by time order or other order. Ensembles are represented as nodes in a graph connected by edges based on the ordering.
9. ihm_ensemble_info: Describe the ensembles obtained as the output of the integrative modeling study, if applicable
10. ihm_ensemble_sub_sample: Details of the sub samples within the ensembles

xv. Localization densities of ensembles

1. ihm_localization_density_files: If localization densities are included in the files connected to the DOI, provide that information in this table.

xvi. Restraints and fitting

Some restraints and fitting data (listed below) are handled in this step. Other restraint data (such as generic distance restraints, chemical crosslinking restraints, hydroxyl radical footprinting restraints and restraints from predicted contacts) that are too tedious to add manually are handled in the next step (Submission step 3: Upload restraint data)

described later.

1. ihm_2dem_class_average_restraint: Details regarding the 2DEM class averages used as restraints in the integrative modeling study
2. ihm_2dem_class_average_fitting: Details regarding the fitting of models to 2DEM images used as restraints
3. ihm_3dem_restraint: Details regarding the 3DEM maps used as restraints in the integrative modeling study
4. ihm_sas_restraint: Details regarding the SAS restraints used as restraints in the integrative modeling study
5. ihm_epr_restraint: Details regarding EPR restraints used in the integrative modeling study

xvii. Chemical descriptors for probes used in experiments

Some experiments (such as FRET) use molecular probes that provide restraint information used in the modeling. The tables below provide the definitions to describe the probes used.

1. ihm_chemical_component_descriptor: Provide chemical information (SMILE strings or INCHI keys) regarding the molecular probes or chemical crosslinking agents used in experiments that provide restraints for the integrative modeling study.
2. ihm_probe_list: List the molecular probes used in the experiments from which spatial restraints are derived (e.g., probes used in FRET experiments).
3. ihm_poly_probe_position: Details of specific residue positions in polymeric entities where probes are covalently attached.
4. ihm_poly_probe_conjugate: Details of probes attached to specific residue positions in polymeric entities
5. ihm_ligand_probe: Details of non-covalently interacting ligands used as probes.

xviii. Descriptions of geometric objects involved in input restraints

Sometimes geometric objects are used as part of a restraint, e.g., a protein bound to a membrane represented as a half-torus. The set of tables listed below provide the definitions to describe geometric objects. The restraints themselves are handled in the next step (Submission step 3: Upload restraint data) described later.

1. ihm_geometric_object_list: List all geometric objects used in the modeling as restraints (e.g., membranes represented as a half-torus or molecules tethered to an axis or a plane).
2. ihm_geometric_object_center: Details of centers of geometric objects
3. ihm_geometric_object_transformation: Details of transformations applied to geometric objects
4. ihm_geometric_object_sphere: Details of spherical geometric objects

5. ihm_geometric_object_torus: Details of torus geometric objects
6. ihm_geometric_object_half_torus: Details of half-torus geometric objects
7. ihm_geometric_object_axis: List any axes used as restraints
8. ihm_geometric_object_plane: List any planes used as restraints

7. Submission Step 3: Upload restraint data

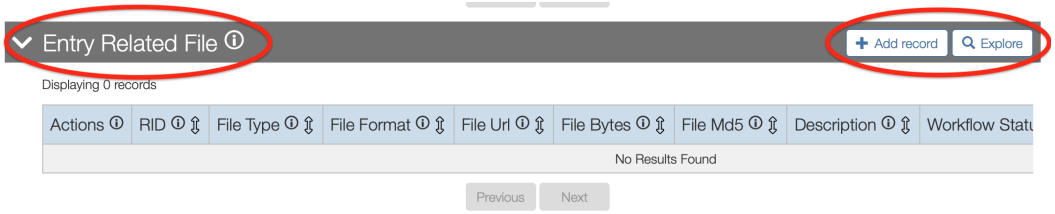
Input restraint data are required for validating integrative models. Based on examples obtained from the integrative modeling community, definitions for restraint data are also included in the IHMCIF dictionary and the Deriva catalog. Since restraint data are typically large, it is not ideal to enter this information manually. Therefore, we allow for restraint data to be uploaded as CSV or TSV files, conforming to the following prerequisites:

a. The first line of the CSV or TSV file should contain the column names separated by commas or tabs as appropriate. The column names should match the column names in the Deriva catalog. CSV/TSV templates can be downloaded from the website and can be used to create the restraint data files for upload.
b. The remaining lines in the file should provide the data pertaining to the column names in the first line, separated by commas.
c. The table “Entry Related File” is used to upload CSV/TSV files containing restraint data pertaining to different restraint tables in the IHMCIF dictionary. Tables for which restraint data can be uploaded as CSV/TSV files are:
    i. ihm_pseudo_site: Details of pseudo sites that may be used in the restraints or model representation
    ii. ihm_crosslink_list: List experimentally determined crosslink distances between molecular entities
    iii. ihm_cross_link_restraint: Details regarding the implementation of crosslinks listed in “ihm_crosslink_list” in the integrative modeling study
    iv. ihm_cross_link_pseudo_site: Details of pseudo sites involved in crosslinks
    v. ihm_cross_link_result: Results of crosslinking restraints used in integrative modeling
    vi. ihm_cross_link_result_parameters: Parameters related to the results of crosslinking restraints
    vii. ihm_predicted_contact_restraint: Details of predicted contact restraints derived from coevolution data
    viii. ihm_hydroxyl_radical_fp_restraint: Details of hydroxyl radical footprinting restraints
    ix. ihm_hdx_restraint: Details of restraint derived from hydrogen-deuterium (H/D) exchange experiments
    x. ihm_feature_list: List polymeric atoms/residues, non-polymers, pseudo sites etc. used to define generic distance restraints

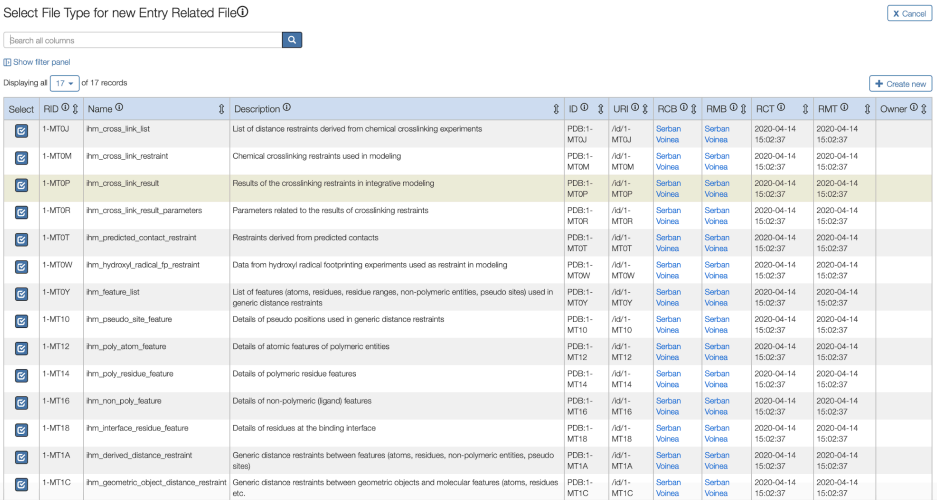
xi. ihm_poly_atom_feature: Details of atomic features in polymeric entities
xii. ihm_poly_residue_feature: Details of residue features in polymeric entities
xiii. ihm_non_poly_feature: Details of non-polymeric features
xiv. ihm_interface_residue_feature: Details of polymeric residue features that are at binding interfaces
xv. ihm_pseudo_site_feature: Details of pseudo site features
xvi. ihm_derived_distance_restraint: Details of the generic distance restraints involving the different kinds of features described above.
xvii. ihm_derived_angle_restraint: Details of the generic angle restraints involving the different kinds of features described above.
xviii. ihm_derived_dihedral_restraint: Details of the generic dihedral restraints involving the different kinds of features described above.
xix. ihm_geometric_object_distance_restraint: Details of distance restraints involving geometric objects described in step 2 and the features described above.

How to upload multiple restraint files corresponding to different tables listed above:

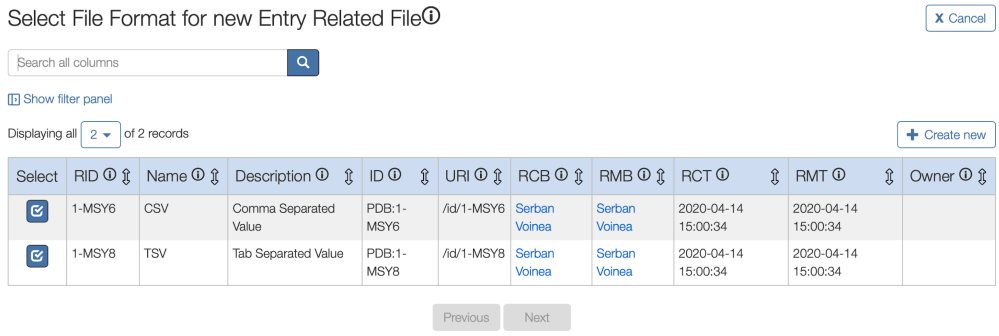
d. Go to the table “Entry Related File” and click on “Add record”



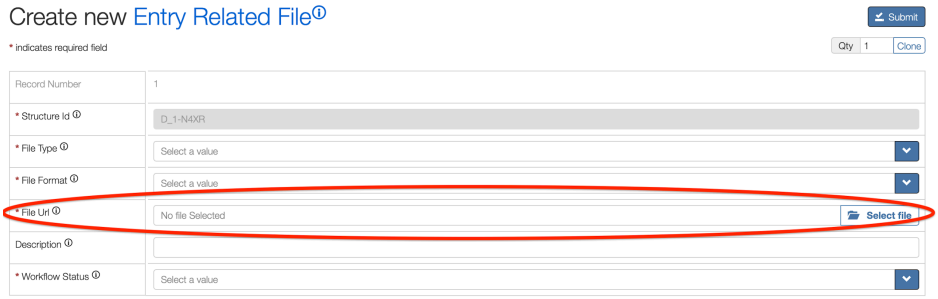
e. Select the “File Type” corresponding to the table for which restraint data is being uploaded



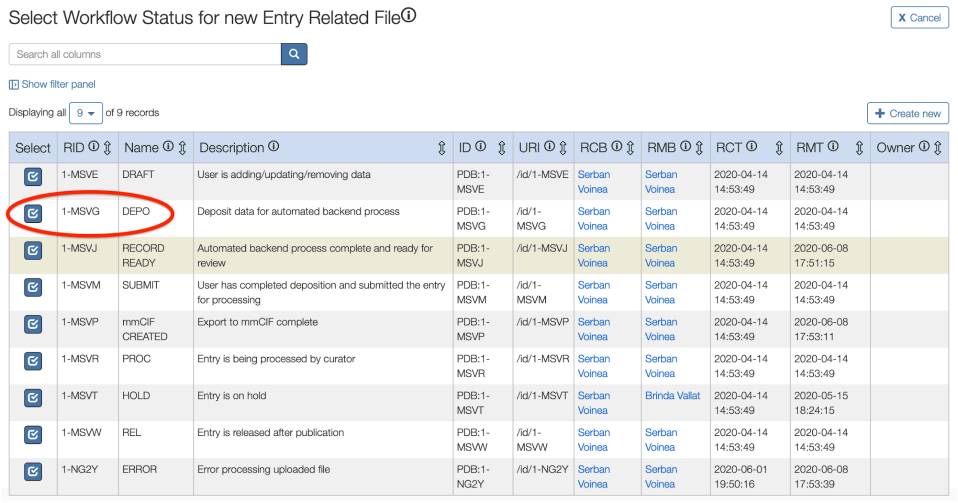
f. Select TSV/CSV file format



g. Upload properly formatted file

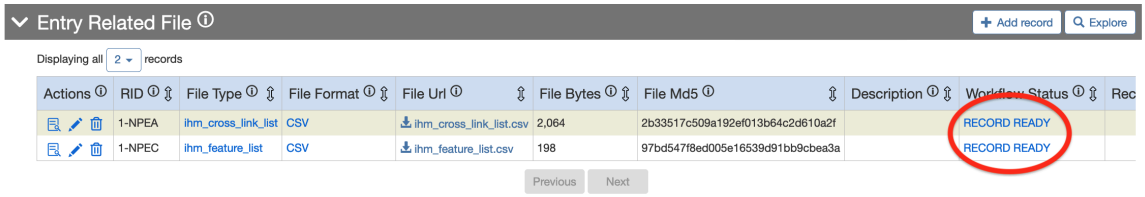


h. Choose “Workflow Status” as “DEPO” and “Submit” the restraint file.



i. Repeat the process for all the relevant restraint tables.

j. The uploaded files will be processed and the corresponding table (chosen in “File Type”) will be populated with the submitted restraint data. If the process is a success, the “Workflow Status” in the “Entry Related File” table corresponding to the particular uploaded file will be updated to “RECORD READY”. If the process fails, the “Workflow Status” in the “Entry Related File” table corresponding to the particular uploaded file will be updated to “ERROR”. The error message will be displayed in the “Record Status Detail” column of the “Entry Related File table”.



k. Once the “Workflow Status” in the “Entry Related File” table has been updated to “RECORD READY”, the data uploaded as TSV/CSV file will be available in the corresponding table and can be visually verified.

8. Tables that are not required to be populated by the user

Some tables are not to be populated by the user: audit_conform, chem_comp_atom, pdbx_entity_poly_na_type, pdbx_entry_details, pdbx_inhibitor_info, pdbx_ion_info, pdbx_protein_info

9. Submission Step 4: Submit entry and wait for mmCIF generation
    a. Once all data are entered / uploaded, the “Workflow Status” column in the “Entry” table needs to be changed to “Submit”.
    b. When the new mmCIF file is generated based on the data provided, the “Workflow Status” column in the “Entry” table will be changed to “mmCIF Created” and the mmCIF file will be available in the “System Generated mmCIF File” table.
    c. If submission of the entry leads to an “ERROR” in the “Workflow Status”, users can try to fix the error based on the information available under “Record Status Detail” and resubmit the entry. Users can contact the PDB-IHM team for any assistance.

10. Accession codes

Accession codes will be provided by PDB-IHM after the submission has been processed to ensure compliance with the IHMCIF dictionary and all the necessary information has been provided. The accession code will be updated in the “entry” table under the “accession code” column.
From August 14 2024, PDB-IHM will start issuing 4-character PDB accession codes as part of the unification process with the PDB archive. Integrative structures processed by PDB-IHM will be made available alongside experimental structures in the PDB archive. PDB-IHM accession codes that have been already issued will be supported and will be available via the `database_2` category in IHMCIF files, but new integrative structures will not be issued PDB-IHM accession codes and only PDB accession codes will be issued going forward.

11. Release process

PDB-IHM entries are updated along with the PDB archive each week on or about Wednesday 00:00 UTC (Coordinated Universal Time) with new releases, modified entries, and updated status information. Updates are prepared on the previous Friday. Citation updates and release requests should be sent to pdb-ihm@mail.wwpdb.org by noon ET on the preceding Thursday to be included in an update; changes made after an update has been packaged will appear with the following update.
The files in the FTP archive have the Thursday timestamp of the internal update packaging.

12. Using the PDB-IHM API for file upload and creating new entries

PDB-IHM accepts depositions via our API. Depositors can use a command line interface (CLI) tool to upload fully compliant IHMCIF files and create new entries for submission.

This tool is particularly useful for bulk uploads and for creating multiple entries belonging to a collection. The tool should only be used for uploading fully compliant IHMCIF files and associated image files. Please reach out to the PDB-IHM team for any assistance.

Instructions for using the CLI:

a. Follow the instructions above to create an account using Globus.
b. Login to https://www.globus.org/ and find the registered globus ID under https://app.globus.org/settings/identities and note it down.
c. In the local workstation, create the following directory: ~/path/to/deriva/{globus_id}/entry
d. Create compliant IHMCIF files as *.cif files and corresponding image files as *.png or *.jpg files. The *.cif and *.png or *.jpg files that correspond to an entry must have the same name with different extensions, e.g., AB-AT.cif and AB-AT.png. Images should be on a transparent background.
e. Put all the *.cif and *.png files belonging to a collection for bulk upload in the ~/path/to/deriva/{globus_id}/entry directory.
f. Software installation (preferably in a virtual environment)
    i. python3 -m venv ./deriva-py-venv
    ii. cd ./deriva-py-venv
    iii. source ./bin/activate
    iv. python3 -m pip install --upgrade pip setuptools wheel
    v. pip3 install deriva
g. Login to the server (only required once)
    i. ./bin/deriva-globus-auth-utils login --refresh --host data.pdb-ihm.org
    ii. Login using the browser
h. Upload files
    i. ./bin/deriva-upload-cli data.pdb-ihm.org ~/path/to/deriva
i. Additional information
    i. First test with one entry (one IHMCIF file and the corresponding image file). If all goes well, upload more.
    ii. Re-uploading files with the same name or md5 will throw an error. In case of errors, please contact the PDB-IHM team for support.
j. Once the entries are created, users can login to https://data.pdb-ihm.org/ to verify the files uploaded and make any necessary changes using the frontend. The “Workflow Status” for entries created without errors will show “Record Ready” at this point.
k. Once the uploaded files and data are verified, the “Workflow Status” column in the “Entry” table must be changed to “Submit”. If the submission goes through, the “Workflow Status” column in the “Entry” table will be changed to “mmCIF Created” and the mmCIF file will be available in the “System Generated mmCIF File” table.

l. If submission of the entry leads to an “ERROR” in the “Workflow Status”, users can try to fix the error based on the information available under “Record Status Detail” and resubmit the entry.