HuBMAP Data Upload Checklist / Tracker

This tool is intended as a supplement to the HuBMAP Data Submission Guide and the Data Ingest Guidance slidedeck. Links to some portals, sites, and other resources are provided herein, but see these other documents for more detail and additional resources regarding each step or section below.

Prerequisites: Complete prior to registering donors and samples or uploading data			
Step	**Comments**		**Status**
1. Complete HuBMAP onboarding process a. Have access to all needed portals:	Applied for membership; can access the HuBMAP Data portal		Com…
i. Globus (How To & FAQs)	Check that your institution ID is linked.		Com…
ii. Write access to team Globus folder	Granted by HuBMAP after confirming that you will be submitting data (*required*).		Com…
iii. Protocols.io	Including HuBMAP Method Development Community (*required*).		Com…
iv. NIH's dbGaP	If you are doing genome sequencing (single cell RNA seq or single cell your raw data would need to be submitted to dbGAP.) Add note to dbGap section. If applicable. sequencing data will be submitted here after your data is published.		N/A
			Not …
b. Submit all preliminary documentation Add more specific directions here (signature)	Consent and data use agreements (add appropriate links here)		Com…
c. All documentation received by HuBMAP			Not …
2. Register your experimental protocol(s)	Via protocols.io; requires a digital object identifier (DOI).		Not …

Clinical items are highlighted like this.

Please CREATE a COPY of this Checklist for your Team's use.

Prerequisites: Complete prior to registering donors and samples or uploading data			
a. Initial setup	Select a lead; create accounts; create a workspace for your team (TMC, TTD, or RTI).		Not …
b. Organize your protocols	Break workflows into modules; create an overall protocol that references others.		Not …
c. Create a New protocol	Many templates are available; pick the DOI option when publishing.		Not …
i. Create & publish a new protocol	Use a template; include “HuBMAP” and team name as a keyword; publish the protocol.		Not …
ii. Modify an existing protocol	Open an existing protocol; use *Edit New Version* or *Edit New Fork*; follow steps above.		Not …

Preparation, Part 1: Registering Donors, Organs, and Samples		
Step	**Comments**	**Status**
1. Register donor via HuBMAP ingest portal	A donor is an individual from which an organ and sample originates.	Not Started
a. Search for the donor	Verify that the donor has not already been registered.	Not Started
b. From the Navigation bar select:	*REGISTER NEW > INDIVIDUAL > Donor*	Not Started
c. Record information about the donor	Do NOT include protected health information about the donor, organs, or specimens.	Not Started
d. Complete registration	Click *Generate ID* to complete the registration process.	Not Started
2. Register an organ or sample.	Prerequisite: The donor associated with the organ or sample has to be registered first.	Not Started
a. From the Navigation bar select:	*REGISTER NEW > INDIVIDUAL > Sample;* (used for both organs and samples)	Not Started
IMPORTANT: A donor and organ *must* be registered first. You need the ID of the organ and the DOI (from protocols.io) for the case selection protocol.		

Clinical items are highlighted like this.

Please CREATE a COPY of this Checklist for your Team’s use.

Preparation, Part 1: Registering Donors, Organs, and Samples		
Step	**Comments**	**Status**
1. Register donor via HuBMAP ingest portal	A donor is an individual from which an organ and sample originates.	Not Started
a. Search for the donor	Verify that the donor has not already been registered.	Not Started
b. From the Navigation bar select:	*REGISTER NEW > INDIVIDUAL > Donor*	Not Started
b. Record information about the sample	*Source ID* field represents the organ or sample's parent, the donor.	Not Started
c. Upload a thumbnail image if the sample is a tissue block.	*Label* the image using this format: *<block_submission_id>.<jpg>.* It should be a stained microscopy image of a tissue section that represents the tissue block, preferably 800x600 pixels, in JPEG or PNG format. See also: Images SOP.	Not Started
d. Is there a 3D reference object for this organ?	Part of the Common Coordinate Framework (CCF) which refers to the entire body. See CCF Portal for listing of supported organs.	
Yes - click *Register Location* No - (no existing CCF)	Register the sample in CCF Coordinate space; see video or download the SOP here. Contact MC-IU if the organ or tissue you are imaging is not represented in the CCF.	Not Started
e. Complete registration	Click *Generate ID* to complete the registration process.	Not Started
f. Download sample metadata template	From GitHub; Send the completed template to HuBMAP Data Curation.	Not Started

Preparation, Part 2: Preparing a Submission Directory and Metadata files		
Step	**Comments**	**Status**
1. Prepare a Submissions Directory	Use for submitting datasets to the HIVE.	
a. Set up the submissions directory	Based on assay-specific directory schema for each assay in Github: (e.g. CODEX Directory structure). These requirements are defined via a collaborative process.	Not Started

Clinical items are highlighted like this.

Please CREATE a COPY of this Checklist for your Team's use.

i. Include these items:	• One assay metadata file per assay type (*assay_metadata.tsv*)	Not Started
	• One contributor's metadata file per dataset (*contributors.tsv*)	Not Started
	• One antibody metadata file per dataset—if applicable—(*antibodies.tsv*)	Not Started
	• One data directory for each dataset	Not Started
NOTE: The next two steps differ for Clinical vs. NON-Clinical assays.		
NON-Clinical (most) assays	Any type of assay other than those specified below as clinical.	
ii. assay-specific directory schema	Access via GitHub link (e.g., CODEX) or *specification and submission* link in the portal.	Not Started
iii. Organize the components of each dataset	(i.e., data and metadata files) in a submission directory according to the required directory structure specified in the GitHub page for the assay (e.g., CODEX).	Not Started
Clinical assays (created in a clinical setting)	*Clinical Assay types include Ultrasound, MRI, OCT, MicroCT, and MicroNanoCT.*	
ii. Create a root directory	Inside this directory place the metadata.tsv and contributors.tsv files.	Not Started
iii. Create a data subdirectory (*datadir*)	Inside the root directory—When all the files are ready, compress the root directory using a utility (e.g. tar or .zip) to reduce its file size for submission.	Not Started
iv. *Extras* directory (optional)	Use for any other files your team wants to include; will not be vetted by the HIVE. ***IMPORTANT*** - Do NOT include TMC-processed data in the *extras* directory.	Not Started
Preparation, Part 2: Preparing a Submission Directory and Metadata files (continued)		
Step	**Comments**	**Status**
2. Create metadata files	Download and complete the *latest version* of each required metadata template.	
a. Assay_metadata.tsv templates	Associated GitHub pages include field name descriptors, format requirements.	Not Started

b. Contributors.tsv template	Lists the names, institutions, and ORCID ID's of contributors to the dataset.	Not Started
c. Antibodies.tsv template	Required for all assays that use antibodies. If *all* datasets in your assay metadata form use *exactly* the same set of antibodies, you may submit just one *antibodies.tsv* file.	Not Started
If datasets in the submission…		
• Use different sets of antibodies OR • The same set with different lots	Each unique antibody set must be listed in a separate *antibody.tsv* with a unique name (*e.g., dataset1_antibodies.tsv , dataset2_antibodies.tsv, etc).* (Different lot numbers for different datasets.)	Not Started
• Custom antibodies	Indicate the batch date (*format: yyyy-mm-dd*) in the lot_number field.	Not Started
• Assay-specific channel_id	The channel_id field in the antibodies TSV is populated with assay-specific information in the following formats:	
➢ CODEX: cycle#_CH#	The cycle/channel information is generated by the Akoya instrument used for CODEX.	Not Started
➢ Cell DIVE: cycle#_CH#	The following 3 assays do not involve cycles:	
➢ Light sheet:	channel_id is the name of the fluorophore tag on the antibody.	
➢ MIBI:	channel_id is the name of the metal tag on the antibody.	
➢ IMC:	channel_id is the name of the metal tag on the antibody.	
Check for RRID (Research Resource Identifiers) registration.	If an RRID has not been assigned for an antibody, register the antibody.	Not Started

Preparation, Part 3: Local Validation

NOTE: TMCs are *required* to run the submission validation locally BEFORE uploading to Globus, but *Clinical assays do not need to complete this step*.

IMPORTANT*:* When a new version of a metadata schema is generated, it will be published on the first day of the *next month*.

Step	Comments	Status

Clinical items are highlighted like this.

Locally validate the submissions directory	Use the *validate_submission.py* code, following the instructions in the README.	Not Started
• Run the validation code	This code validates the submission directory content against the appropriate assay-specific metadata & data directory schema.	Not Started
• Correct any errors; repeat until no errors occur.	A successful validation run will end with a "No errors!" message. Contact the HuBMAP help desk for assistance with *persistent* errors.	Not Started
Preparation, Part 4: Creating a Data Upload directory / folder (NON-Clinical Assays only)		
Submit metadata to the HIVE	*Prerequisite*: The submission directory has been successfully validated locally.	Not Started
1. Go to the HuBMAP ingest portal		Not Started
2. From the Navigation bar select:	*REGISTER NEW > BULK > Data;* (Begins the data upload process).	Not Started
3. Complete the required fields.	Select a title that both you and the HIVE can use to recognize the data upload.	Not Started
4. Click *Create* to create a folder for your datasets.	Creates a folder in Globus where you upload your dataset files.	Not Started
5. Need to add or modify files?	Follow the highlighted *Globus* link.	Not Started
a. *Save*:	To make changes to the *Title* or *Description* fields use the *Save* button.	Not Started
b. *Submit*:	Click *Submit* once all data has been uploaded to Globus. Contact HuBMAP help desk with the details of the data upload if further processing by the HIVE is needed.	Not Started

Preparation, Part 4: Creating a Data Upload directory / folder (Clinical Assays only)		
Step	**Comments**	**Status**
1. Email the HuBMAP help desk when ready to upload.	• Be sure to use an institutional email address (e.g. *name@institution.edu*). • The Help desk relays this request for space on the protected endpoint to the	Not Started

Clinical items are highlighted like this.

Please CREATE a COPY of this Checklist for your Team's use.

	Pitt team. ● The Pitt team provides the assay team with the *protected data endpoint.*	
2. Upload your data as compressed / zipped folders, using Globus.	***IMPORTANT:*** This location is distinct from where NON-Clinical HIVE data is uploaded.	Not Started
Data Upload: Submitting Donor Data		
1. Email the HuBMAP help desk to…	*Prerequisite*: Your submission donors and samples have been registered.	Not Started
a. Request setup of a donor directory	***NOTE:*** Donor privacy is a top priority when HuBMAP receives data from the community.	Not Started
b. HIVE creates a Globus directory	"Please upload HuBMAP donor data only here - this site is HIPAA protected so if de-identification errors are made we will correct them - please associate all content with the applicable donor HuBMAP ID - those are of the format: HBM111.XXXX.111..."	Not Started
c. You will receive email notification. *The email will state:*		Not Started
2. Upload donor data / clinical information matched to the correct HuBMAP IDs	Once all of the data has been uploaded, reply to the original Help Desk email / ticket.	Not Started
a. IEC begins processing your donor data	To ensure capture of as much data as possible for potential querying by scientists.	Not Started
b. Use the Globus endpoint to upload	For better performance and support, given the size of HuBMAP data files.	Not Started
See the Globus FAQ regarding Globus Connect and Endpoints, HuBMAP's data-related policies, or contact the HuBMAP help desk with any additional questions.		

Data Processing: Ingestion, HIVE Validation, and Processing		
Step	**Comments**	**Status**
Prerequisites: ● All files uploaded into Globus folder	All data files have been uploaded into the team's Globus folder.	Not Started

Clinical items are highlighted like this.

Please CREATE a COPY of this Checklist for your Team's use.

• Each dataset has a directory	Each dataset has a corresponding directory and row in an associated *metadata.tsv* file.	Not Started
• Each metadata.tsv file has one row for each dataset	Multiple tsv files can be included, but at least one is *required*.	Not Started
NOTE: Multiple sets of data can be uploaded and the same data set *could* be uploaded more than once. Each data upload is assigned a UUID (universally unique identifier), but not a version number, until it has completed the entire data ingest, validation, and approval process and is published.		
1. Email the HuBMAP help desk	Once all the files associated with an upload have been uploaded. This lets them know that the upload is ready for ingestion (not currently an automated process.) ***IMPORTANT:*** Include the *root path(s)* of the specific data upload(s) in the email.	Not Started
2. The HIVE will extract each data upload	And corresponding data and register them as individual datasets.	Not Started
3. The HIVE will process the data and metadata to be ingested.	If ingestion fails or if additional information is needed, the HIVE will contact the data provider using the ticketing system.	Not Started
NOTE: For Clinical assay data, the Pitt team will de-identify this data and submit the scrubbed data to the HIVE. The provider of the data will need to review and approve the release of the de-identified data in the publish step.		
Ingestion statuses:	Displayed in the ingestion portal as a data set is processed.	
• New	Data upload registered; Globus directory where files are uploaded created.	Not Started
• Submitted	Data upload submitted for validation and processing by the HIVE.	Not Started
• Processing	The data upload is currently being processed and is not editable.	Not Started
Data Processing: Ingestion, HIVE Validation, and Processing (continued)		
Step	**Comments**	**Status**
• Valid	The data upload has been successfully validated.	Not Started

• Reorganized (into datasets)	The data has been reorganized into the datasets referenced in the metadata.tsv file(s).	Not Started
• Invalid	This data upload did not pass validation.	
• Error	An (unspecified) error occurred during HIVE processing.	
4. Pipeline processing (currently):	The HIVE will process certain assays by standardized pipelines (where applicable).	
a. CODEX	Cyclic immunofluorescence imaging: Pipeline uses Cytokit + SPRM	Not Started
b. ATAC-seq	(Including sc-, sn-, and bulk variants): Pipeline uses SnapTools + SnapATAC	Not Started
c. RNA-seq	(Including sc-,sn-,bulk): Pipeline uses Salmon + Scanpy + scVelo	Not Started
d. Other Imaging data: SPRM		Not Started
Additional pipelines will be added over time. Pipelines are available for download use by others, including HuBMAP TMCs. ***NOTE**:* Generally, TMCs are not involved with pipelines, but may be contacted if an error occurs.		
Review: Final Review, Data Approval, and Publication		
• Final review of data. • ***NOTE**:* For pipeline-processed data:	All HuBMAP data contributors will perform a final review of their data. For data processed using a HIVE pipeline, the pipeline *must complete* without throwing a programmatic error prior to the final data review.	Not Started Not Started
• Data and metadata pass validation	The HIVE contacts the data provider (PI or PM) with instructions on how to access the pre-release data for their approval before it is made public.	Not Started
Data Upload: Submission to dbGaP (Sequence data)		
Step	**Comments**	**Status**
Submit data to dbGaP after publication • Joint process - data provider & HIVE	***NOTE:*** This process can take up to three months from the beginning of the process, through the final release of the data.	

1. HIVE contacts data provider PI & PM	After the release of data in the HuBMAP portal to begin the process.	Not Started
2. Register a study, create bioproject	The Data Provider PI and Genomic Program Administrator (GPA) register the study in the dbGaP Submission System and create a bioproject for the study.	Not Started
3. Designate data submitters and add them to the dbGaP study	• The data provider PI or PM identifies a team member to serve in this capacity. This person can contact the HIVE via the HuBMAP help desk. • The HIVE identifies a HIVE data submitter to work with the team member on the dbGap data submission. • The Data Provider PI adds the data submitters as submitters for the dbGaP study on the NCBI dbGaP submission portal. ***NOTE:*** *This is distinct from the dbGaP Submission system identified in step 2, above.*	Not Started Not Started Not Started
4. Submission Portal Questionnaire	The Data Provider PI or their designated data submitter completes this on the NCBI dbGaP submission portal for the registered study (step 2 above).	Not Started
5. Required forms for submission (see list below)	Following the dbGaP instructions, the designated data submitter and HIVE data submitter download the dbGaP Submission Guide Templates and work together to complete these forms. *Gather and complete ALL files BEFORE submitting.*	Not Started
• Study Config	To be completed by the designated data submitter	Not Started
• Subject Consent DS & DD	To be completed by the HIVE data submitter (DS = Dataset, DD = Data Dictionary)	Not Started
• Subject Sample Mapping (SSM) DS & DD	To be completed by the HIVE data submitter	Not Started
• Sample Attributes DS & DD	To be completed by the HIVE data submitter	Not Started
Data Upload: Submission to dbGaP (Sequence data) - continued		
Step	**Comments**	**Status**
NOTE: Depending on the answers provided to the Submission Portal Questionnaire, the designated data submitter may also need to complete the following files:		

• Mapping Study Samples	To Other NCBI databases (e.g. Sample GEO Mapping DS and DD).	Not Started
• Subject Phenotypes DS & DD	(DS = Dataset, DD = Data Dictionary)	Not Started
• Pedigree DS and DD		Not Started
• Study Documents	consent forms, protocols, etc.	Not Started
6. Reviews checklist. submits forms	The HIVE data submitter reviews to ensure that the Phenotype Datasets and Data dictionary files pass the dbGaP quality control tests; then submits all required forms.	Not Started
7. dbGaP curator contacts submitters	Phenotype curator contacts the submitters once the above information has been loaded into dbGaP and entities have received NCBI BioSample and SRA IDs.	Not Started
8. Upload sequencing metadata	The HIVE data submitter follows directions here to upload sequencing metadata and here to submit raw sequence reads working with an SRA curator, to a protected area of SRA.	Not Started
9. SRA processes the sequence data and metadata	SRA notifies dbGaP and the submitters.	Not Started
10. Human sequence data distributed	Through Authorized Access upon dbGaP release of the study. ***IMPORTANT:*** This process can take an additional 6-8 weeks.	Not Started