Instructions for the Submission of Data and Metadata to HCA for Atlas Assembly

Version: v1.1 - issued 25.02.25 (refer amendments to the HCA Executive Office)

Last updated: 25.02.25

Overview of the ingestion process

1. Purpose of this document

The following instructions are for scientists wanting to contribute molecular data to the Human Cell Atlas (HCA). This document describes the file formats accepted, how and where data needs to be ingested and the required metadata.

The Human Cell Atlas (HCA) is creating comprehensive reference maps of all human cells as a basis for understanding human health and disease. Cellular maps are constructed by integrating many datasets together requiring a standardised metadata schema.

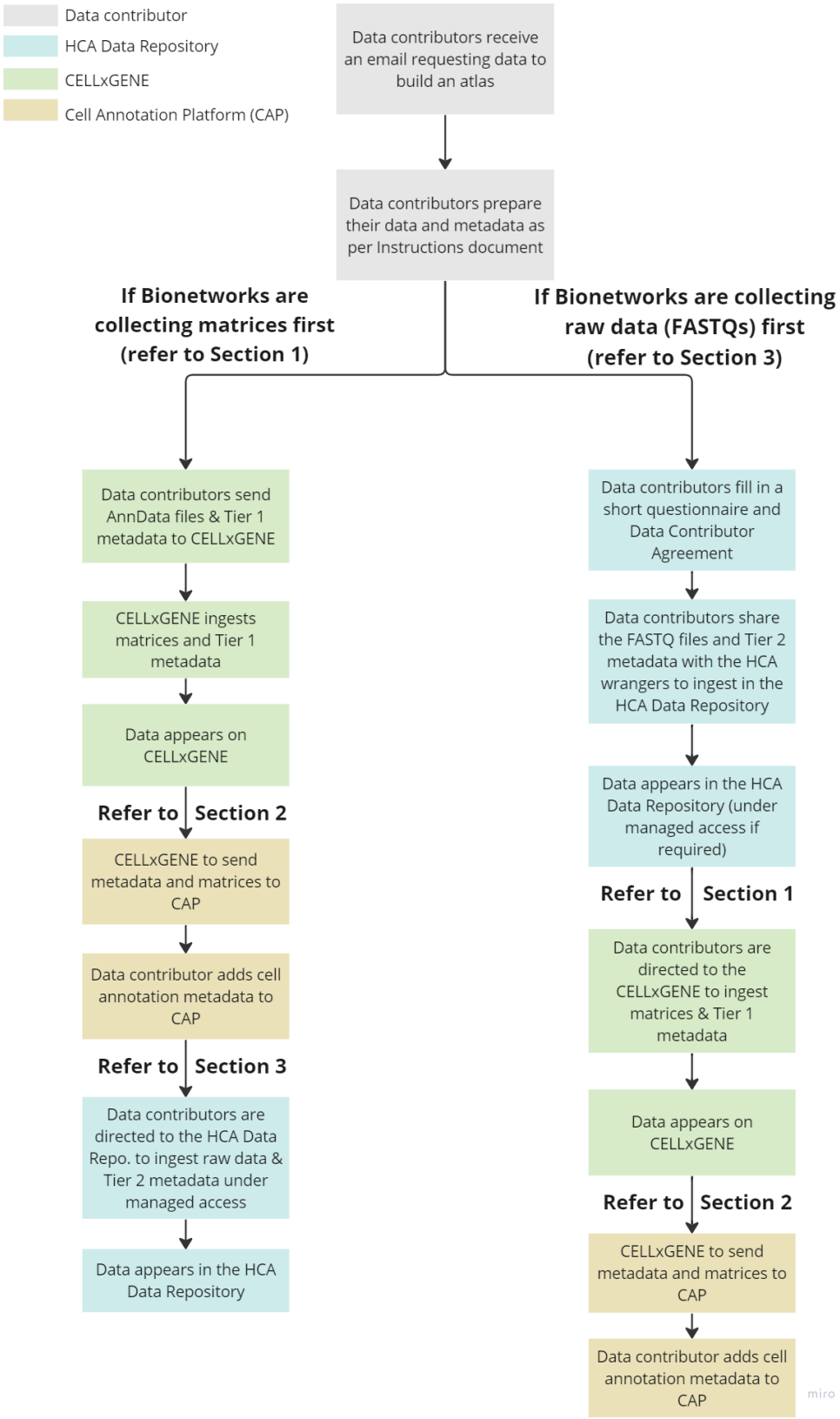
2. What needs to be submitted to be part of an atlas?

All data contributors must provide the following data and metadata for inclusion in an atlas:

- Gene expression matrices and corresponding Tier 1 metadata (Section 1)
    - Tier 1 metadata are fields are those required to assemble an atlas
- Cell annotation metadata (Section 2)
- Raw data files and Tier 2 metadata (managed access) (Section 3)
    - Tier 2 metadata fields are all additional fields helpful for biological analysis

While all data is important, the submission of the gene expression matrices and the Tier 1 metadata (Section 1) are prerequisites to start building an atlas.

3. Flow diagram of ingestion process



4. Data policies and protection measures

4.1. Unpublished data:

- Contributors of unpublished data are encouraged to make their data publicly available immediately via the HCA Data Repository, in accordance with the HCA Data Release Policy.
- If the data contributors decline to do so, they must, at minimum, agree to make their data publicly available via the HCA Data Repository (for raw sequence data and detailed Tier 2 metadata) and Chan Zuckerberg CELLxGENE (for gene expression matrices, Tier 1 metadata, and broad demographic metadata) as soon as the Atlas that incorporates it is published in a peer-reviewed scientific journal. We invite you to discuss any concerns you may have with your Bionetwork Coordinators.
- If you are contributing unpublished data, you may request that your data be "embargoed" in a private HCA data repository until the organ/tissue/system atlas you have contributed to is published. At that point, your data will be made available publicly via the HCA Data Explorer and linked to the corresponding Atlas page on the HCA Data Portal. Data contributors will be notified before their data is made public.
- Access to embargoed data will be restricted to HCA Data Wranglers, the corresponding Atlas integration team, and anyone else specified by the data contributor.
- For more information refer to HCA's Unpublished Data Policy

4.2. Managed access data:

HCA has a managed access service. Data submitted to the HCA Data Repository (Section 3) can be protected by managed access if elected to by the data contributor. Access to managed access data is controlled through the HCA Data Access Committee. Scientists seeking access to these data must submit a data access request which is reviewed by the HCA Data Access Committee.

4.3. Submission of metadata to other repositories

Data contributors need to notify HCA if biological metadata is submitted to other data repositories.

Section 1. Instructions for the submission of gene expression matrices and corresponding Tier 1 metadata fields required to commence integration

1. Overview of the Tier 1 metadata

Description	Tier 1 metadata fields provide the foundational information used to build tissue and organ atlases.
Priority	Submission of the Tier 1 fields is a prerequisite for atlas building and should be prioritised.
Rationale	Tier 1 metadata fields are required to: ● Help identify samples or datasets included in integration ● Help identify and explain technical batch effects ● Help qualify or disqualify datasets for inclusion in atlases Understanding the factors that can cause batch effects is vital to ensure that when the datasets are combined into an atlas, they have not been over-corrected (i.e., masking true biological variation between cells) or under-corrected (e.g., resulting in the same cell types appearing as distinct from one another).
Stored by	CELLxGENE and the Cell Annotation Platform (CAP)
Examples	obs: cell_enrichment; sample_collection_method; CL_term uns: publication_doi

2. File formats required for submission

Data contributors wishing to submit datasets to HCA for integration into atlases are required to submit an AnnData file. AnnData files combine matrices and metadata into a single file, thus metadata should be **captured on a per cell basis**.

An AnnData file has several components including:

- uns (Dataset metadata), which describe the dataset as a whole
- X (Matrix layers), which describe the data required for different assays
- obs (Cell metadata), which describe each cell in the dataset
- obsm (Embeddings), which describe each embedding in the dataset
- obsp, which describe pairwise annotation of observations
- var and raw.var (Gene metadata), which describe each gene in the dataset
- varm, which describe multidimensional annotation of variables/features
- varp, which describe pairwise annotation of variables/features

File format specifications

- **Dataset-level metadata in uns**:
  - title
  - study_pi
  - batch_condition
  - default_embedding
  - comments

- **Data in .X and raw.X:**
  - raw counts are required
  - normalized counts are strongly recommended
  - raw counts should be in raw.X if normalized counts are in .X
  - if there is no normalized matrix, raw counts should be in .X

- **Cell metadata in obs**: Metadata fields are to be captured on a **per cell** basis.
  - protocol_url
  - donor_id
  - sample_id
  - institute
  - sample_collection_site
  - sample_collection_relative_time_point
  - library_id
  - library_id_repository
  - author_batch_notes
  - organism_ontology_term_id
  - manner_of_death
  - sample_source
  - sex_ontology_term_id
  - sample_collection_method
  - sample_preservation_method
  - suspension_type
  - cell_enrichment
  - cell_viability_percentage
  - cell_number_loaded
  - sample_collection_year
  - assay_ontology_term_id
  - library_preparation_batch
  - library_sequencing_run
  - sequenced_fragment
  - sequencing_platform
  - is_primary_data
  - reference_genome
  - gene_annotation_version
  - alignment_software

- tissue_type
- sampled_site_condition
- tissue_ontology_term_id
- tissue_free_text
- intron_inclusion
- author_cell_type
- cell_type_ontology_term

In addition to the HCA obs fields above, there are an additional three fields that are required for submission into CELLxGENE. These fields are not part of the HCA Tier 1 metadata fields.

- disease_ontology_term_id
- development_stage_ontology_term_id
- self_reported_ethnicity_ontology_term_id

- **Embeddings in obsm**:
  - One or more two-dimensional embeddings, prefixed with 'X_'

- **Features in var & raw.var (if present)**:
  - index is Ensembl ID
  - preference is that gene have not been filtered in order to maximise future data integration efforts

Should you require support creating an AnnData file or converting your file from another single cell file format click here or reach out to cellxgene@chanzuckerberg.com and mention that you are contributing to the HCA.

3. List of fields

The list of the Tier 1 metadata fields (including descriptions and examples) can be found in the Appendix 1. A google sheet template of all fields that vary on the sample level (i.e., excluding cell type fields) is also available. Please note, for submission, the fields must be included in an AnnData object (not submitted in spreadsheet format).

HCA has expanded the CELLxGENE metadata schema by adding fields capturing information essential for integration.

4. Instructions for ingestion

AnnData files will be stored and are accessible on CELLxGENE Discover. To submit files, please follow the process for submission:

1. Please reach out to the curation team at cellxgene@chanzuckerberg.com with an email containing the following information.
   - Title

- Description
- Contact: name and email
- Publication/preprint DOI: The publication digital object identifier (doi) for the protocol. If no pre-print nor publication exists, please write 'not applicable'.
- URLs: any additional URLs for related data or resources, such as GEO or protocols.io - can be added later
- Consortia (i.e. HCA)

2. The team confirms acceptance of your data.
3. You prepare your AnnData file and send the file to cellxgene@chanzuckerberg.com.
4. The team will upload your dataset to a private collection where you can review.
5. The team will make your dataset either: public or private. Private datasets need to be shared with the integration team for inclusion in the integrated object.

5. Schema versioning

HCA will continue to align with the most up-to-date CELLxGENE schema.

6. Matrices already stored in CELLxGENE

If you have matrices that are already stored in CELLxGENE you will only need to add additional fields required for HCA atlas assembly. Please contact CELLxGENE and provide a table with the new fields detailed and the team can (in most cases) merge the additional fields into the existing AnnData file. In some complex cases, the AnnData file will need to be resubmitted.

The outstanding fields are:

Uns:

- study_pi
- comments

Obs: Captured on a **per cell** basis.

- protocol_url
- sample_id
- institute
- sample_collection_site
- sample_collection_relative_time_point
- library_id
- library_id_repository
- author_batch_notes
- manner_of_death
- sample_source

- sample_collection_method
- sampled_site_condition
- tissue_free_text
- sample_preservation_method
- cell_enrichment
- cell_viability_percentage
- cell_number_loaded
- sample_collection_year
- library_preparation_batch
- library_sequencing_run
- sequenced_fragment
- sequencing_platform
- is_primary_data
- reference_genome
- gene_annotation_version
- alignment_software
- intron_inclusion
- author_cell_type

7. Unpublished data

If you are contributing unpublished data, your data will be kept in a private space until the organ/tissue/system atlas you have contributed to is published. At this point, you will be prompted to make your data available publicly.

Note: As a tool that is open and available for use by the broader scientific community, CELLxGENE can also ingest datasets that are not being used for atlas assembly. To do this, contributors should visit the website.

Section 2. Instructions for the submission of cell annotation metadata

1. Overview of cell annotation metadata

Description	These metadata fields relate to the naming of cells and will be stored on the Cell Annotation Platform (CAP). CAP is a platform that allows individuals or groups to work on cell annotations in a private space before deciding to 'publish' the annotations for public viewing. At the time of initial submission, a dataset does not need to have a finalised list of all cell annotations - that is one of the functions of CAP - it is a workspace to define cell annotations over time. You will find that the cell annotation metadata requests detailed cell names (including ontology terms and synonyms), as well as the names of the overarching 'parent' populations (called 'cell categories'). This helps to align cell names across datasets and create a hierarchy. We highly recommend using the CAP user interface to input cell annotations as it is easy to use and has helpful suggestions and prompts.
Rationale	Cell annotation fields are collected in order to: ● Find similar cell annotations between different cell annotation sets ● Define the similarities of the molecular signatures used to define individual "cell types" ● Define coarse cell annotations agreeing between dataset (could be through defining broad cellular hierarchy); this will help with the process of data integration and help QC the integrated object by e.g., assessing if similarly annotated cell types co-localize in integrated embedding; ● Identifying rarer cell types across datasets that could be used to QC you data integration assessing that rarer cell types are not lost in the integration process (in integration SOP, referred to as "seed annotation") ● Create an "agreement/consensus" between cell annotations

Stored by	Cell Annotation Platform (CAP)
Examples	CL_term; marker_genes_evidence

2. List of fields

The list of cell annotation metadata fields can be found in the Appendix 2.

If data contributors are eager to start collecting cell annotation metadata fields in advance while waiting for your data matrices and Tier 1 metadata to be validated and ingested into CAP, the following Google sheets template can be used.

3. Instructions for ingestion

Data contributors will be notified once their datasets have been validated and upload to CAP has been initiated. Instructions will be provided on how to create a profile on CAP and enter cell annotation metadata - this is easily done using the CAP user interface.

1. In order to accept the invite and edit the draft workspace, data curators will need to create a profile on CAP with details such as name, contact email and academic institution.
2. Data contributors will receive an email from CAP inviting them to a draft workspace on the CAP website containing information already provided to the HCA wranglers during the submission of data matrices as well as the Tier 1 validated datasets.
3. Data contributors will need to complete the cell annotation metadata fields for all cell labels using the CAP platform's user interface. This can be done by an individual or by a group working within the private workspace.
    a. Collaborators can be invited to contribute to the workspace by workspace owners and admins by clicking on Collaborators.
4. Once the required fields have been completed, the draft publication will be submitted to review by CAP data curators and then published on the CAP website for others to view.
    a. Should new annotation sets be added or edited, a new version of the annotations can be published.

Note: As a tool that is open and available for use by the broader scientific community,

CAP can also ingest datasets that are not being used for atlas assembly. To do this, contributors should visit the CAP website and sign up or log in. Detailed instructions on how to submit, annotate and then publish annotations are available here.

Section 3. Instructions for the submission of raw data files and managed access Tier 2 metadata

1. Overview of Tier 2 metadata

Description	There are two groups of fields that are captured under the umbrella of Tier 2 metadata: - Fields that capture information that is not strictly required for the integration of datasets but is highly valuable for downstream biological analysis (these fields may differ per Biological Network) - Fields requiring managed access protection (i.e. fields that contain potential identifiers)
Rationale	These fields are separated from the Tier 1 fields for two reasons. First, these fields may take longer to wrangle and we do not want to delay the teams building the integrated objects. Second, many of these fields require managed access protection for some data contributors and are therefore stored in a separate system with the necessary controls.
Stored by	HCA Data Repository (formally 'DCP')
Examples	Ethnicity_self_reported; BMI

2. File formats required for submission

FASTQ files are accepted.

Metadata should be submitted to the HCA wranglers via email. The metadata is captured in template spreadsheets provided to the data contributors by the HCA wranglers. To offer more protection to managed access datasets, the metadata spreadsheet will not be shared via email but will be uploaded to a secure folder instead.

3. List of fields

The metadata fields will be tailored specifically to each Biological Network. A tailored list of fields will be distributed to data contributors when the process commences and will be highlighted in the metadata spreadsheet template.

The lists of fields are being defined by members of the Biological Network and authors contributing datasets to the atlas.

4. Instructions for ingestion

To submit raw data (FASTQ files) and Tier 2 metadata, please follow the process for submission:

1. All data contributors will be emailed to complete two tasks:
    a. Complete a short questionnaire to: (1) tailor the metadata template to the experimental design, and (2) determine whether the dataset needs to be embargoed.
    b. Complete the mandatory HCA 'Data Contributor Agreement' form which is required to ensure all data accepted into the HCA has the requisite protection measures therefore adhering to local data protection laws and regulations. Upon completion you will automatically be sent a copy of the form for your records and another copy will be sent to HCA for record keeping.
2. Upon receipt of the questionnaire and Data Contributor Agreement, data contributors will be:
    a. Provided with access to a secure folder where the FASTQ files should be uploaded.
    b. Sent a metadata spreadsheet listing all relevant fields required for meaningful biological enquiry into the atlas. Data contributors are asked to fill in the metadata spreadsheet and upload it to the secure access folder.
3. The HCA Data Repository wrangling team reviews the metadata spreadsheet and uploads the dataset on the HCA Data Repository.
4. Embargoed datasets will remain private until the atlas they are a part of is published. Under embargo, they will only be shared with the integration team building the atlas.
   If a contributor wishes to publish their embargoed dataset at an earlier date they can do so by contacting the wrangling team at wrangler-team@data.humancellatlas.org.

Appendix 1: Tier 1 Metadata

Collection metadata:

The following fields need to be included in the submission email.

- Title
- Description
- Contact: name and email
- Publication/preprint DOI: The publication digital object identifier (doi) for the protocol. If no pre-print nor publication exists, please write 'not applicable'.
- URLs: any additional URLs for related data or resources, such as GEO or protocols.io - can be added later
- Consortia (i.e. HCA)

AnnData file:

Each dataset needs the following information added to a single h5ad (AnnData 0.8) format file.

X (Matrix layers):

The data stored in the X data matrix is the data that is viewable in CELLxGENE Explorer. CELLxGENE does not impose any additional constraints on the X data matrix.

In any layer, if a matrix has 50% or more values that are zeros, it is STRONGLY RECOMMENDED that the matrix be encoded as a scipy.sparse.csr_matrix.

CELLxGENE's matrix layer requirements are tailored to optimize data reuse. Because each assay has different characteristics, the requirements differ by assay type. In general, CELLxGENE requires submission of "raw" data suitable for computational reuse when a standard raw matrix format exists for an assay. It is STRONGLY RECOMMENDED to also include a "normalized" matrix with processed values ready for data analysis and suitable for visualization in CELLxGENE Explorer. So that CELLxGENE's data can be provided in download formats suitable for both R and Python, the schema imposes the following requirements:

- All matrix layers MUST have the same shape, and have the same cell labels and gene labels.
- Because it is impractical to retain all barcodes in raw and normalized matrices, any cell filtering MUST be applied to both. By contrast, those wishing to reuse datasets require access to raw gene expression values, so genes SHOULD NOT

be filtered from either dataset. Summarizing, any cell barcodes that are removed from the data MUST be filtered from both raw and normalized matrices and genes SHOULD NOT be filtered from the raw matrix.

- Any genes that publishers wish to filter from the normalized matrix MAY have their values replaced by zeros and MUST be flagged in the column feature_is_filtered of var, which will mask them from exploration.
- Additional layers provided at author discretion MAY be stored using author-selected keys, but MUST have the same cells and genes as other layers. It is STRONGLY RECOMMENDED that these layers have names that accurately summarize what the numbers in the layer represent (e.g. "counts_per_million", "SCTransform_normalized", or "RNA_velocity_unspliced").

The following table describes the matrix data and layers requirements that are assay-specific. If an entry in the table is empty, the schema does not have any other requirements on data in those layers beyond the ones listed above.

Assay	"raw" required?	"raw" location	"normalized" required?	"normalized" location
scRNA-seq (UMI, e.g. 10x v3)	REQUIRED. Values MUST be de-duplicated molecule counts. Each cell MUST contain at least one non-zero value. All non-zero values MUST be positive integers stored as numpy.float32.	AnnData.raw.X unless no "normalized" is provided, then AnnData.X	STRONGLY RECOMMENDED	AnnData.X
scRNA-seq (non-UMI, e.g. SS2)	REQUIRED. Values MUST be one of read counts (e.g. FeatureCounts) or estimated fragments (e.g. output of RSEM). Each cell MUST contain at least one non-zero value. All non-zero values MUST be positive integers stored as numpy.float32.	AnnData.raw.X unless no "normalized" is provided, then AnnData.X	STRONGLY RECOMMENDED	AnnData.X

Accessibility (e.g. ATAC-seq, mC-seq)	NOT REQUIRED		REQUIRED	AnnData.X

Uns:

Field name	**title**
Description	This text describes and differentiates the dataset from other datasets in the same collection. It is strongly recommended that each dataset title in a collection is unique and does not depend on other metadata such as a different assay to disambiguate it from other datasets in the collection.
Values	N/A
Required	MUST
Examples	Cells of the adult human heart collection is "All — Cells of the adult human heart".

Field name	**study_pi**
Description	Principal Investigator(s) leading the study where the data is/was used.
Required	MUST
Examples	Sarah,A,Teichmann

Field name	**batch_condition**
Description	*Note: Name of the covariate that confers the dominant batch effect in the data as judged by the data contributor. The name provided here should be the label by which this covariate is stored in the AnnData object.* Values must refer to cell metadata keys in obs. Together, these keys define the batches that a normalisation or integration algorithm should be aware of. For example if "patient" and "seqBatch" are keys of vectors of cell metadata, either ["patient"], ["seqBatch"], or ["patient", "seqBatch"] are valid values.
Required	RECOMMENDED
Examples	-

Field name	**default_embedding**
Description	The value must match a key to an embedding in obsm for the embedding to display by default in CELLxGENE Explorer.
Required	RECOMMENDED
Examples	-

Field name	**comments**
Description	Other technical or experimental covariates that could affect the quality or batch of the sample. Must not contain identifiers. This field is designed to capture potential challenges for data integration not captured elsewhere.

Required	RECOMMENDED
Examples	-

obsm (Embeddings):

The size of the ndarray stored for a key in obsm MUST NOT be zero.

To display a dataset in CELLxGENE Explorer, Curators MUST annotate one or more embeddings of at least two-dimensions (e.g. tSNE, UMAP, PCA, spatial coordinates) as numpy.ndarrays in obsm.

var and raw.var (Gene Metadata):

var and raw.var are both of type pandas.DataFrame.
Curators MUST annotate the following columns in the var dataframe and if present, the raw.var dataframe.

index of pandas.DataFrame

Key	index of pandas.DataFrame
Annotator	Curator
Value	str. If the feature is a gene then this MUST be an ENSEMBL term. If the feature is a RNA Spike-In Control Mix then this MUST be an ERCC Spike-In identifier (e.g. "ERCC-0003"). The index of the pandas.DataFrame MUST contain unique identifiers for features. If present, the index of raw.var MUST be identical to the index of var.

Curators MUST annotate the following column only in the var dataframe. This column MUST NOT be present in raw.var:

feature_is_filtered

Key	feature_is_filtered
Annotator	Curator
Value	bool. This MUST be True if the feature was filtered out in the normalized matrix (X) but is present in the raw matrix (raw.X). The value for all cells of the given feature in the normalized matrix MUST be 0. Otherwise, this MUST be False.

Curators MUST NOT annotate the following columns in the var dataframe and if present, the raw.var dataframe.

Obs:

Field name	sample_id
Description	Identification number of the sample. This is the fundamental unit of sampling the tissue (the specimen taken from the subject), which can be the same as the 'donor_ID', but is often different if multiple samples are taken from the same subject. Note: this is NOT a unit of multiplexing of donor samples, which should be stored in "library".
Values	N/A
Required	MUST
Rationale	Fundamental unit of sampling of the tissue.
Examples	SC24; SC25; SC28

Field name	donor_id
Description	This must be free-text that identifies a unique individual that data were derived from.
Values	It is strongly recommended that this identifier be designed so that it is unique to: a given individual within the collection of datasets that includes this dataset, and a given individual across all collections in CELLxGENE Discover. It is strongly recommended that "pooled" be used for observations from a sample of multiple individuals that were not confidently assigned to a single individual through demultiplexing. It is strongly recommended that "unknown" ONLY be used for observations in a dataset when it is not known which observations are from the same individual.
Required	MUST
Rationale	Fundamental unit of biological variation of the data
Examples	CR_donor_1; MM_donor_1; LR_donor_2

Field name	protocol_url
Description	The protocols.io URL (if none exists, please use the BioRxiv URL) for the full experimental protocol; or if multiple protocols exist please list them e.g. sample preparation protocol / sequencing protocol.
Values	N/A
Required	RECOMMENDED
Rationale	Useful to look up protocol data that can provide insight on batch effects. As protocols can sometimes apply to a subset of the study, we capture this at a sample level. This information may not always be available.
Examples	https://www.biorxiv.org/conte nt/early/2017/09/24/193219

Field name	institute

Description	Institution where the samples were processed.
Values	N/A
Required	MUST
Rationale	To be able to link to other studies from the same institution as sometimes samples from different labs in the same institute are processed via similar core facilities. Thus batch effects may be smaller for datasets from the same institute even if other factors differ.
Examples	EMBL-EBI; Genome Institute of Singapore

Field name	**sample_collection_site**
Description	The pseudonymised name of the site where the sample was collected.
Values	It is strongly recommended that this identifier be designed so that it is unique to a given site within the collection of datasets that includes this site (for example, the labels 'site1', 'site2' may appear in other datasets thus rendering them indistinguishable).
Required	RECOMMENDED
Rationale	To understand whether the collection site contributes to batch effects
Examples	AIDA_site_1; AIDA_site_2

Field name	**sample_collection_relative_time_point**
Description	Time point when the sample was collected. This field is only needed if multiple samples from the same subject are available and collected at different time points. Sample collection dates (e.g. 23/09/22) cannot be used due to patient data protection, only relative time points should be used here (e.g. day3).
Values	N/A
Required	RECOMMENDED
Rationale	Explains variability in the data between samples from the same subject
Examples	sampleX_day1

Field name	**library_id**
Description	The unique ID that is used to track libraries in the investigator's institution (should align with the publication).
Values	N/A
Required	MUST
Rationale	A way to track the unit of data generation. This should include sample pooling
Examples	A24; NK_healthy_001

Field name	**library_id_repository**
Description	The unique ID used to track libraries from one of the following public data repositories: EGAX*, GSM*, SRX*, ERX*, DRX, HRX, CRX
Values	N/A
Required	RECOMMENDED
Rationale	Links a dataset back to the source from which it was ingested, optional only if this is the same as the library_id.

Examples	GSM1684095

Field name	**author_batch_notes**
Description	Encoding of author knowledge on any further information related to likely batch effects.
Values	N/A
Required	RECOMMENDED
Rationale	Space for author intuition of batch effects in their dataset
Examples	Batch run by different personnel on different days

Field name	**organism_ontology_term_id**
Description	The name given to the type of organism, collected in NCBITaxon:0000 format.
Values	"NCBITaxon:9606" for Homo sapiens or "NCBITaxon:10090" for Mus musculus.
Required	MUST
Rationale	Strong biological effect that needs to be considered for batch covariate selection
Examples	NCBITaxon:9606; NCBITaxon:10090

Field name	**manner_of_death**
Description	Manner of death classification based on the Hardy Scale or 'unknown' or 'not applicable': • Category 1 = Violent and fast death Deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 min. • Category 2 = Fast death of natural causes -Sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hr (with sudden death from a myocardial infarction as a model cause of death for this category) • Category 3 = Intermediate death - Death after a terminal phase of 1 to 24 hrs (not classifiable as 2 or 4); patients who were ill but death was unexpected • Category 4 = Slow death - Death after a long illness, with a terminal phase longer than 1 day (commonly cancer or chronic pulmonary disease); deaths that are not unexpected • Category 0 =Ventilator Case - All cases on a ventilator immediately before death • Unknown = The cause of death is unknown • Not applicable = Subject is alive [Please leave this field as blank for embryonic/fetal tissue]
Values	1; 2; 3; 4; 0; unknown; not applicable
Required	MUST
Rationale	Manner of death can affect cellular profiles.
Examples	1; 2; 3; 4; 0; unknown; not applicable

Field name	**sample_source**
Description	The study subgroup that the participant belongs to. This indicates whether the participant was a surgical donor (this includes patients providing blood samples or biopsies), a postmortem donor, or an organ donor.

Values	surgical donor; postmortem donor; living organ donor
Required	MUST
Rationale	The source of the sample (whether the sample comes from alive subject; an organ donor; or deceased subject) can result in different cellular profiles and hence batch effects.
Examples	surgical donor; postmortem donor

Field name	**sex_ontology_term_id**
Description	Reported sex of the donor.
Values	This must be a child of PATO:0001894 for phenotypic sex or "unknown" if unavailable.
Required	MUST
Rationale	Likely biological effect. Need to know if we have a balanced dataset or if sex is collinear with the dataset.
Examples	PATO:0000383 for female, PATO:0000384 for male

Field name	**sample_collection_method**
Description	The method the sample was physically obtained from the donor.
Values	brush; scraping; biopsy; surgical resection; blood draw; body fluid; other
Required	MUST
Rationale	Main contributor to batch effects
Examples	biopsy; brush; surgical resection

Field name	**tissue_type**
Description	Whether the tissue is "tissue", "organoid", or "cell culture".
Values	tissue; organoid; cell culture
Required	MUST
Rationale	Source of batch effect & dataset exclusion criteria
Examples	tissue; organoid; cell culture

Field name	**sampled_site_condition**
Description	Whether the site is considered healthy, diseased or adjacent to disease.
Values	healthy; diseased; adjacent
Required	MUST
Rationale	Main contributor to batch effects
Examples	healthy; diseased; adjacent

Field name	**tissue_ontology_term_id**
Description	The detailed anatomical location of the sample, please provide a specific UBERON term.
Values	If tissue_type is "tissue" or "organoid", this must be the most accurate child of UBERON:0001062 for anatomical entity. If tissue_type is "cell culture" this must follow the requirements for cell_type_ontology_term_id.

Required	MUST
Rationale	Major biological effect that needs to be assessed for sufficient coverage in the atlas datasets.
Examples	UBERON:0001828; UBERON:0000966

Field name	**tissue_free_text**
Description	The detailed anatomical location of the sample - this does not have to tie to an ontology term.
Values	N/A
Required	RECOMMENDED
Rationale	To help the integration team understand the anatomical location of the sample, specifically to solve the problem when the UBERON ontology terms are insufficiently precise.
Examples	terminal ileum

Field name	**sample_preservation_method**
Description	Indicating if tissue was frozen, or not, at any point before library preparation.
Values	ambient temperature; cut slide; fresh; frozen at -70C; frozen at -80C; frozen at -150C; frozen in liquid nitrogen; frozen in vapor phase; paraffin block; RNAlater at 4C; RNAlater at 25C; RNAlater at -20C; other
Required	MUST
Rationale	Main contributor to batch effects
Examples	fresh; frozen at -70C

Field name	**suspension_type**
Description	Specifies whether the sample contains single cells or single nuclei data.
Values	This must be "cell", "nucleus", or "na". This must be the correct type for the corresponding assay: ● 10x transcription profiling [EFO:0030080] and its children = "cell" or "nucleus" ● ATAC-seq [EFO:0007045] and its children = "nucleus" ● BD Rhapsody Whole Transcriptome Analysis [EFO:0700003] = "cell" ● BD Rhapsody Targeted mRNA [EFO:0700004] = "cell" ● CEL-seq2 [EFO:0010010] = "cell" or "nucleus" ● CITE-seq [EFO:0009294] and its children = "cell" ● DroNc-seq [EFO:0008720] = "nucleus" ● Drop-seq [EFO:0008722] = "cell" or "nucleus" ● GEXSCOPE technology [EFO:0700011] = "cell" or "nucleus" ● inDrop [EFO:0008780] = "cell" or "nucleus"
Required	MUST
Rationale	Major source of batch effect & dataset exclusion criteria
Examples	cell; nucleus; na

Field name	**cell_enrichment**
Description	Specifies the cell types targeted for enrichment or depletion beyond the selection of live cells.

Values	This must be a Cell Ontology (CL) term (http://www.ebi.ac.uk/ols4/ontologies/cl). For cells that are enriched, list the CL code followed by a "+". For cells that were depleted, list the CL code followed by a "-". If no enrichment or depletion occurred, please use 'na' (not applicable)
Required	MUST
Rationale	If cell lineages were filtered, this may be a dataset exclusion criterion
Examples	CL:0000057+; na

Field name	**cell_viability_percentage**
Description	If measured, per sample cell viability before library preparation (as a percentage).
Values	N/A
Required	RECOMMENDED
Rationale	Is a measure of sample quality that could be used to explain outlier samples
Examples	88; 95; 93.5

Field name	**cell_number_loaded**
Description	Estimated number of cells loaded for library construction.
Values	N/A
Required	RECOMMENDED
Rationale	Can explain the number of doublets found in samples
Examples	5000; 4000

Field name	**sample_collection_year**
Description	Year of sample collection. Should not be detailed further(to exact month and day), to prevent identifiability.
Values	N/A
Required	RECOMMENDED
Rationale	May explain whether a dataset was separated into smaller batches.
Examples	2018

Field name	**assay_ontology_term_id**
Description	Platform used for single cell library construction.

Values	This must be an EFO term and either: • "EFO:0002772" for assay by molecule or preferably its most accurate child • "EFO:0010183" for single cell library construction or preferably its most accurate child • An assay based on 10X Genomics products should either be "EFO:0008995" for 10x technology or preferably its most accurate child. An assay based on SMART (Switching Mechanism at the 5' end of the RNA Template) or SMARTer technology SHOULD either be "EFO:0010184" for Smart-like or preferably its most accurate child. Recommended: 10x 3' v2 "EFO:0009899" 10x 3' v3 "EFO:0009922" 10x 5' v1 "EFO:0011025" 10x 5' v2 "EFO:0009900" Smart-seq2 "EFO:0008931" Visium Spatial Gene Expression "EFO:0010961"
Required	MUST
Rationale	Major source of batch effect and dataset filtering criterion
Examples	EFO:0009922

Field name	**library_preparation_batch**
Description	Indicating which samples' libraries were prepared in the same chip/plate/etc., e.g. batch1, batch2.
Values	N/A
Required	MUST
Rationale	Sample preparation is a major source of batch effects.
Examples	batch01; batch02

Field name	**library_sequencing_run**
Description	The identifier (or accession number) that indicates which samples' libraries were sequenced in the same run.
Values	N/A
Required	MUST
Rationale	Library sequencing is a major source of batch effects
Examples	run1; NV0087

Field name	**sequenced_fragment**
Description	Which part of the RNA transcript was targeted for sequencing.
Values	3 prime tag; 5 prime tag; probe-based; full length
Required	MUST
Rationale	May be a source of batch effect that has to be tested
Examples	3 prime tag; full length

Field name	sequencing_platform
Description	Platform used for sequencing.
Values	"subClassOf" : ["EFO:0002699"] - https://www.ebi.ac.uk/ols/ontologies/efo/terms?iri=http%3A%2F%2Fwww.ebi.ac.uk%2Fefo%2FEFO_0002699
Required	RECOMMENDED
Rationale	This captures potential strand hopping which may cause data quality issues
Examples	EFO:0008563

Field name	is_primary_data
Description	This must be True if this is the canonical instance of this cellular observation and False if not. This is commonly False for meta-analyses reusing data or for secondary views of data.
Values	true; false
Required	MUST
Rationale	This helps to ensure samples are not used twice.
Examples	true; false

Field name	reference_genome
Description	Reference genome used for alignment.
Values	GRCh38; GRCh37; GRCm39; GRCm38; GRCm37; not applicable
Required	MUST
Rationale	Possible source of batch effect and confounder for some biological analysis
Examples	GRCh38; GRCh37

Field name	gene_annotation_version
Description	Ensembl release version accession number. Some common codes include: GRCh38.p12 = GCF_000001405.38 GRCh38.p13 = GCF_000001405.39 GRCh38.p14 = GCF_000001405.40
Values	http://www.ensembl.org/info/website/archives/index.html) or NCBI/RefSeq
Required	MUST
Rationale	Possible source of batch effect and confounder for some biological analysis
Examples	v110; GCF_000001405.40

Field name	alignment_software
Description	Protocol used for alignment analysis, please specify which version was used e.g. cell ranger 2.0, 2.1.1 etc.
Values	N/A
Required	MUST
Rationale	Affects which cells are filtered per dataset, and which reads (introns and exons or only exons) are counted as part of the reported transcriptome. This can convey batch effects.

Examples	cell ranger 3.0.1; kallisto bustools; GSNAP

Field name	**intron_inclusion**
Description	Were introns included during read counting in the alignment process?
Values	yes; no
Required	RECOMMENDED
Rationale	Affects the number of reads per cell called in a sample
Examples	yes; no

Field name	**author_cell_type**
Description	Encoding of author intuition of cellular annotation in the dataset.
Values	N/A
Required	RECOMMENDED
Rationale	Encoding of author intuition of cellular annotation in their dataset.
Examples	Goblet cell; microglia

Field name	**cell_type_ontology_term_id**
Description	Cell Ontology (CL) term.
Values	This must be a Cell Ontology (CL) term (http://www.ebi.ac.uk/ols4/ontologies/cl). If no appropriate high-level term can be found or the cell type is unknown, then it is strongly recommended to use 'unknown'. The following terms must not be used: "CL:0000255" for eukaryotic cell; "CL:0000257" for Eumycetozoan cell; "CL:0000548" for animal cell
Required	MUST
Rationale	Encoding of cell type to help alignment with other datasets.
Examples	CL:0001204

Additional fields required for submission into CELLxGENE:

The following three fields are special cases as they are required for submission into CELLxGENE but are not part of HCA's formal Tier 1 metadata schema.

For contributors of data within the jurisdiction of GDPR (or another authority or regulatory body with like standards), HCA requires contributors to adhere to the specific practices when submitting this information to CELLxGENE. These requirements can be found in each specification below.

Field name	**disease_ontology_term_id**
Description	Disease, if expected to impact the sample.
Values	This must be a MONDO term or "PATO:0000461" for normal or healthy. **Requirements for data contributors adhering to GDPR or like standards**: In the case

	of disease, HCA requests that you submit a higher order ontology term - this is especially important in the case of rare disease.
Required	MUST
Rationale	CELLxGENE core schema
Examples	MONDO:0005385; PATO:0000461

Field name	**self_reported_ethnicity_ontology_term_id**
Description	Self reported ethnicity of the subject.
Values	If organism_ontolology_term_id is "NCBITaxon:9606" for *Homo sapiens*, this must be a HANCESTRO term or "unknown". Otherwise, for all organisms, this must be “na”. **Requirements for data contributors adhering to GDPR or like standards**: HCA will be collecting ethnicity data as part of HCA’s Tier 2 metadata that is protected by managed access, therefore please put 'unknown' for this field.
Required	MUST
Rationale	CELLxGENE core schema
Examples	unknown; HANCESTRO:0008

Field name	**development_stage_ontology_term_id**
Description	Age of the subject.
Values	If organism_ontolology_term_id is "NCBITaxon:9606" for *Homo sapiens*, this should be an HsapDv term. If organism_ontolology_term_id is "NCBITaxon:10090" for *Mus musculus*, this should be an HsapDv term. **Requirements for data contributors adhering to GDPR or like standards**: HCA requests that you do not submit year-specific terms. For convenience, below are some broader age bracket ontology terms: ● Embryonic stage = A term from the set of Carnegie stages 1-23 = (up to 8 weeks after conception; e.g. HsapDv:0000003) ● Fetal development = A term from the set of 9 to 38 week post-fertilization human stages = (9 weeks after conception and before birth; e.g. HsapDv:0000046) ● Post natal = ○ Years 0-14 HsapDv:0000264 ○ Years 15-19 HsapDv:0000268 ○ Years 20-29 HsapDv:0000237 ○ Years 30-39 HsapDv:0000238 ○ Years 40-49 HsapDv:0000239 ○ Years 50-59 HsapDv:0000240 ○ Years 60-69 HsapDv:0000241 ○ Years 70-79 HsapDv:0000242 ○ Years 80-89 HsapDv:0000243
Required	MUST
Rationale	CELLxGENE core schema
Examples	HsapDv:0000237; unknown

Appendix 2: Cell Annotation Metadata

Github link to CAP schema:

[single-cell-curation/schema at main · chanzuckerberg/single-cell-curation · GitHub]

Obs (Cell metadata):

NOTE: The index for obs must be the Cell barcode names.

Field name	Clustering
	Users may OPTIONALLY include a single field for clustering within AnnData files, or multiple fields denoting clustering, e.g. different clustering algorithms, multiple resolutions of clustering, etc.
	We therefore REQUIRE that clustering is clearly denoted within the AnnData file if it contains clustering fields.
	ScanPy has set an AnnData community standard of defining the *.obs value by the type of algorithm. e.g. the function scanpy.tl.louvain (documented here) by default saves the clustering as anndata.obs['louvain']. Similarly, leiden (documented here) is often encoded as anndata.obs['leiden'].
format	The column name is 'cluster', 'leiden', 'louvain' or 'cluster + _ + [ALGORITHM_TYPE] + _ + [SUFFIX]' whereby [ALGORITHM_TYPE] and [SUFFIX] are OPTIONAL. ● 'cluster', 'leiden' or 'louvain': MUST be used to denote clustering in AnnData.obs ● [ALGORITHM]: Denotes the algorithm used, e.g. be either 'leiden' or 'louvain'. OPTIONAL. ● [SUFFIX]: Denotes a descriptive tag informative enough for third-party users; used to distinguish between multiple clusterings. OPTIONAL.
dtype	category
value	Integer of cluster label.
source	file
required for publication on CAP	no
example column name	'cluster_leiden' or 'cluster_leiden_broad' or 'louvain' or 'leiden'
example value	'0' or '1' or '2'

Cell Annotation Metadata:

Field name	[cellannotation_setname]
note	The string specified by the user for [cellannotation_setname] will be used as the pandas DataFrame column name (key) to encode the following cell annotation metadata columns in *.obs. NOTE: A dataset may have multiple sets of cell annotations each with a corresponding set of cell annotation metadata, e.g. 'cell_type' and 'broadclustering_celltype'.
format	The column name is the string [cellannotation_setname] and the values are the strings of cell_label. Refer to the fields cellannotation_setname and cell_label in the JSON Schema.
dtype	string
value	Any free-text term which the author uses to annotate cells, the preferred cell label name used by the author.
source	file or UI
required for publication on CAP	yes
example	'HBC2' or 'rod bipolar'

Field name	[cellannotation_setname]--cell_fullname
format	The column name is the value [cellannotation_setname] concatenated with the string 'cell_fullname' and two hyphens, i.e. [cellannotation_setname] + '--' + 'cell_fullname' For example, if the user specified the cell annotation as broad_cells1, then the name of the column in the pandas DataFrame will be broad_cells1--cell_fullname.
dtype	string
value	The full-length name for the biological entity listed in [cellannotation_setname] by the author.
source	file or UI
required for publication on CAP	yes
example	'rod bipolar'

Field name	[cellannotation_setname]--cell_ontology_exists
format	The column name is the string prefix [cellannotation_setname]-- concatenated with the string value cell_ontology_exists, i.e. [cellannotation_setname] + '--' + 'cell_ontology_exists'
dtype	boolean
value	Boolean value in Python (either True or False).
source	file or UI
required for publication on CAP	yes

example	'True'

Field name	**[cellannotation_setname]--cell_ontology_term_id**
format	The column name is the value [cellannotation_setname] concatenated with the string 'cell_ontology_term_id' and two hyphens, i.e. [cellannotation_setname] + '--' + 'cell_ontology_term_id'
dtype	string
value	This MUST be a term from either the Cell Ontology or from some ontology that extends it by classifying cell types under terms from the Cell Ontology e.g. the Provisional Cell Ontology or the Drosophila Anatomy Ontology (DAO).
source	file or UI
required for publication on CAP	yes
example	'CL:0000751'

Field name	**[cellannotation_setname]--cell_ontology_term**
format	The column name is the value [cellannotation_setname] concatenated with the string 'cell_ontology_term' and two hyphens, i.e. [cellannotation_setname] + '--' + 'cell_ontology_term'
dtype	string
value	The human-readable name associated with the cell_ontology_term_id.
source	file or UI
required for publication on CAP	yes
example	'rod bipolar cell'

Field name	**[cellannotation_setname]--rationale**
format	The column name is the value [cellannotation_setname] concatenated with the string 'rationale' and two hyphens, i.e. [cellannotation_setname] + '--' + 'rationale'
dtype	string
value	The free-text rationale which users provide as justification/evidence for their cell annotations.
source	file or UI
required for publication on CAP	yes
example	'This cell was annotated with [blank] given the canonical markers in the field [X], [Y], [Z]. We noticed [X] and [Y] running differential expression.'

Field name	**[cellannotation_setname]--rationale_dois**
format	The column name is the value [cellannotation_setname] concatenated with the string 'rationale_dois' and two hyphens, i.e. [cellannotation_setname] + '--' + 'rationale_dois'
dtype	string
value	Comma-separated string of valid publication DOIs cited by the author to support or

	provide justification/evidence/context for cell_label.
source	file or UI
required for publication on CAP	no
example	'10.1038/s41587-022-01468-y, 10.1038/s41556-021-00787-7, 10.1038/s41586-021-03465-8'

Field name	**[cellannotation_setname]--marker_gene_evidence**
format	The column name is the value [cellannotation_setname] concatenated with the string 'marker_gene_evidence' and two hyphens, i.e. [cellannotation_setname] + '--' + 'marker_gene_evidence'
dtype	string
value	Comma-separated string of HGNC gene names explicitly used as evidence for this cell annotation. Each gene MUST be included in the matrix of the AnnData/Seurat file.
source	file or UI
required for publication on CAP	yes
example	'TP53, KRAS, BRCA1'

Field name	**[cellannotation_setname]--canonical_marker_genes**
format	The column name is the value [cellannotation_setname] concatenated with the string 'canonical_marker_genes' and two hyphens, i.e. [cellannotation_setname] + '--' + 'canonical_marker_genes'
dtype	string
value	Comma-separated string of gene names considered to be canonical markers for the biological entity used in the cell annotation.
source	file or UI
required for publication on CAP	yes
example	'GATA3, CD3D, CD3E'

Field name	**[cellannotation_setname]--synonyms**
format	The column name is the value [cellannotation_setname] concatenated with the string 'synonyms' and two hyphens, i.e. [cellannotation_setname] + '--' + 'synonyms'
dtype	string
value	Comma-separated string of synonyms for values in [cellannotation_setname]. Abbreviations are acceptable.
source	file or UI
required for publication on CAP	yes
example	'neuroglial cell, glial cell, neuroglia' or 'amacrine cell' or 'FMB cell'

The following fields relate to the ‘category’. Categories are classes of cells of which the cell type is a

subclass. For example, “T lymphocyte” could be a category of “CD8+ T cell”. Although categories will be broader than the selected cell type, it is recommended to select the most specific category that fits your cell type.

Field name	**[cellannotation_setname]--category_fullname**
format	The column name is the string prefix [cellannotation_setname]-- concatenated with the string value category_fullname, i.e. [cellannotation_setname] + '--' + 'category_fullname'. This MUST be the full-length name for the biological entity, not an abbreviation.
dtype	string
value	A single value of the category/parent term for the cell label value in [cellannotation_setname].
source	file or UI
required for publication on CAP	yes
example	'ON-bipolar cell'

Field name	**[cellannotation_setname]--category_cell_ontology_exists**
format	The column name is the string prefix [cellannotation_setname]-- concatenated with the string value category_cell_ontology_exists, i.e. [cellannotation_setname] + '--' + 'category_cell_ontology_exists'
dtype	boolean
value	Boolean value in Python (either True or False).
source	file or UI
required for publication on CAP	yes
example	'True'

Field name	**[cellannotation_setname]--category_cell_ontology_term_id**
format	The column name is the value [cellannotation_setname] concatenated with the string 'synonyms' and two hyphens, i.e. [cellannotation_setname] + '--' + 'category_cell_ontology_term_id'
index	Cell barcode names
dtype	string
value	The ID from either the Cell Ontology or from some ontology that extends it by classifying cell types under terms from the Cell Ontology.
source	file or UI
required for publication on CAP	yes
example	'CL:0000749'

Field name	**[cellannotation_setname]--category_cell_ontology_term**
format	The column name is the string prefix [cellannotation_setname]-- concatenated with the string value category_cell_ontology_term, i.e. [cellannotation_setname] + '--' + 'category_cell_ontology_term'

dtype	string
value	The human-readable name assigned to the value of 'category_cell_ontology_term_id'.
source	file or UI
required for publication on CAP	yes
example	'ON-bipolar cell'
Field name	**[cellannotation_setname]--cell_ontology_assessment**
format	The column name is the string prefix [cellannotation_setname]-- concatenated with the string value cell_ontology_assessment, i.e. [cellannotation_setname] + '--' + 'cell_ontology_assessment'
dtype	string
value	Free-text field for researchers to express disagreements with any aspect of the Cell Ontology for this cell annotation.
source	file or UI
required for publication on CAP	no
example	'Amacrine cell should have four child terms: glycinergic, GABAergic, GABAergic Glycinergic amacrine cells and non-GABAergic non-glycinergic amacrine cells; which then contain their corresponding child terms'

uns (Dataset metadata):

NOTE: Python dictionary within the uns dictionary, with the key the string [cellannotation_setname]

Field name	**cellannotation_metadata**
key	[cellannotation_set]--metadata
type	python dictionary
value	The rest of the dictionary as defined below.
source	file or UI
required for publication on CAP	yes
example	'{[cellannotation_set]--metadata:{'annotation_method':'algorithmic'...}}'
Field name	**cellannotation_setdescription**
key	'description'
type	string
value	Description of the cellannotation_set created. This is free-text for collaborators and third-parties to understand the context/background for the creation of this cell annotation set. We STRONGLY recommend this field be descriptive for other scientists unfamiliar

	with this project to understand why this set of cell annotations exist.
source	file or UI
required for publication on CAP	yes
example	'Cell annotations based on resolution broad clustering using the Leiden algorithm.'

Field name	**annotation_method**
key	'annotation_method'
type	string
value	'algorithmic', 'manual', or 'both' NOTE: If 'algorithmic' or 'both', more details are required. If 'manual', the values in the following 'algorithm_' and 'reference_' fields will be 'NA'.
source	file or UI
required for publication on CAP	yes
example	'algorithmic' or 'manual' or 'both'

Field name	**algorithm_name**
key	'algorithm_name'
type	string
value	The name of the algorithm used.
source	file or UI
required for publication on CAP	yes
example	'scArches' or if 'manual' then 'NA'

Field name	**algorithm_version**
key	'algorithm_version'
type	string
value	The string of the algorithm's version, which is typically in the format '[MAJOR].[MINOR]', but other versioning systems are permitted based on the algorithm's versioning.
source	file or UI
required for publication on CAP	yes
example	'0.5.9' or if 'manual' then 'NA'

Field name	**algorithm_repo_url**
key	'algorithm_repo_url'
type	string
value	The string of the URL of the version control repository associated with the algorithm used (if applicable). It MUST be a string of a valid URL.

source	file or UI
required for publication on CAP	yes
example	'https://github.com/theislab/scarches' or if 'manual' then 'NA'

Field name	**reference_location**
key	'reference_location'
type	string
value	The string of the URL pointing to the reference dataset.
source	file or UI
required for publication on CAP	no
example	'https://figshare.com/projects/Tabula_Muris_Senis/64982' or if 'manual' then 'NA'

Field name	**reference_description**
key	'reference_description'
type	string
value	Free-text description of the reference used for automated annotation for this cell annotation set. Users are welcome to write out context which may be useful for other researchers.
source	file or UI
required for publication on CAP	no
example	'Tabula Muris Senis: a single cell transcriptomic atlas across the life span of Mus musculus which includes data from 18 tissues and organs.' or if 'manual' then 'NA'

Appendix 3: Tier 2 Metadata

These fields are being defined and tailored specifically to each Biological Network.The list of fields will be shared with data contributors when they are due to commence ingestion of FASTQ files into the HCA Data Repository.