Preparing imaging data and metadata for the HTAN DCC

Adam Taylor
Senior Scientist, Sage Bionetworks

SageBionetworks

A prelude…

Courtesy of Milen Nikolov @ Sage

Enabling FAIR data requires communication

We want to know if these data are useful for our problem.

Data users

Enabling FAIR data requires communication

We have a dataset with some files. How should we submit it?

Data contributors

We want to know if these data are useful for our problem.

Data users

Enabling FAIR data requires communication

We have a dataset with some files. How should we submit it?

Data contributors

How do we know what problem is relevant for your data? **Describe** data well.

Data coordinators

We want to know if these data are useful for our problem.

Data users

Enabling FAIR data requires communication

We have a dataset with some files. How should we submit it?

Data contributors

How should we describe our data? We need this out **soon**!

How do we know what problem is relevant for your data? **Describe** data well.

Data coordinators

We want to know if these data are useful for our problem.

Data users

Enabling FAIR data requires communication

We have a dataset with some files. How should we submit it?

Data contributors

How should we describe our data? We need this out **soon**!

How do we know what problem is relevant for your data? **Describe** data well.

Data coordinators

Contextualize your data. That way it'd be interoperable too!

We want to know if these data are useful for our problem.

Data users

Enabling FAIR data requires communication

We have a dataset with some files. How should we submit it?

Data contributors

Say WHAaAT now?!

How should we describe our data? We need this out **soon**!

How do we know what problem is relevant for your data? **Describe** data well.

Data coordinators

Contextualize your data. That way it'd be interoperable too!

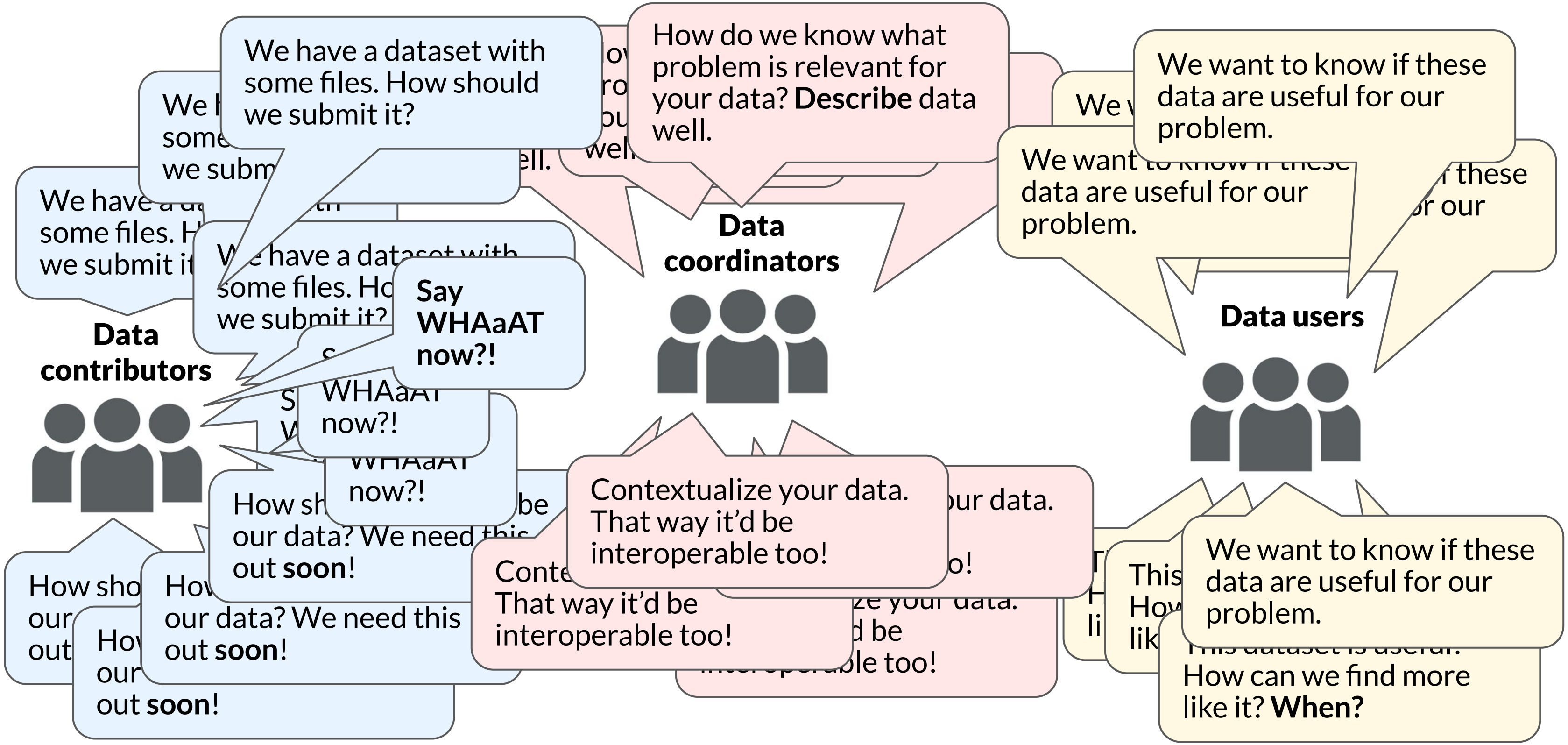
We want to know if these data are useful for our problem.

Data users

This dataset is useful! How can we find more like it? **When?**

Enabling FAIR data requires COMMUNICATION

at SCALE...



We have a dataset with some files. How should we submit it?

Say **WHAaAT now?!**

Data contributors

How sh… our data? We need this out **soon**!

Ho… our data? We need this out **soon**!

How do we know what problem is relevant for your data? **Describe** data well.

Data coordinators

Contextualize your data. That way it'd be interoperable too!

We want to know if these data are useful for our problem.

Data users

We want to know if these data are useful for our problem.

How can we find more like it? **When?**


James


Milen


Adam


Ashley


Elivra


Vésteinn


Ethan


Rochelle


Niki


Ino


Bill


Clarisse


Dave


Dar’ya

Aaron Lisman

Bruno Grande

Mialy DeFelice

Gianna Jordan

Onur Sumer

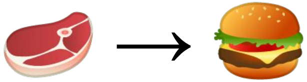
Xiang Li

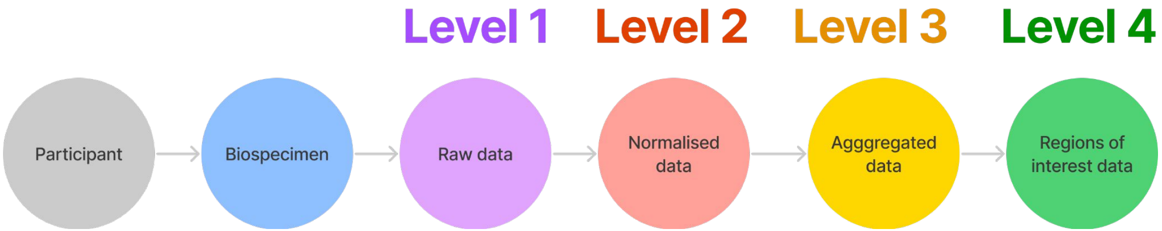
Pierrette Lo

Tiarah Jeudi

Framework for data and metadata management

Data levels:
(Inspired by TGCA)

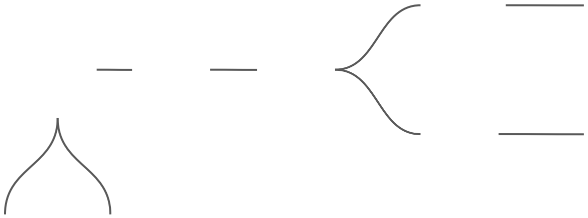



Level 1
Level 2
Level 3
Level 4

Participant → Biospecimen → Raw data → Normalised data → Agggregated data → Regions of interest data

Community-driven
Request for Comments

Unified schema providing
validation and provenance



☐ **HTAN IDs:** HTA9_04_3329

Modern, extensible tools for
schema development and
data/metadata management

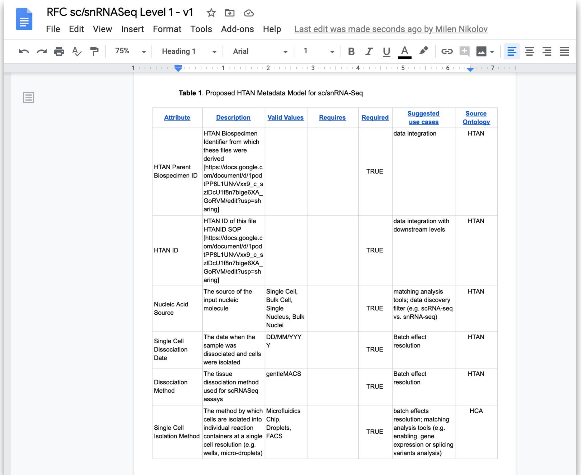
schematic python client

Data Curator App

Shiny
Flask
JSON-LD
schema.org

Data model development

HTAN Schema
758 attributes, 36 components,
>6000 nodes (incl valid values)



RFC sc/snRNASeq Level 1 - v1

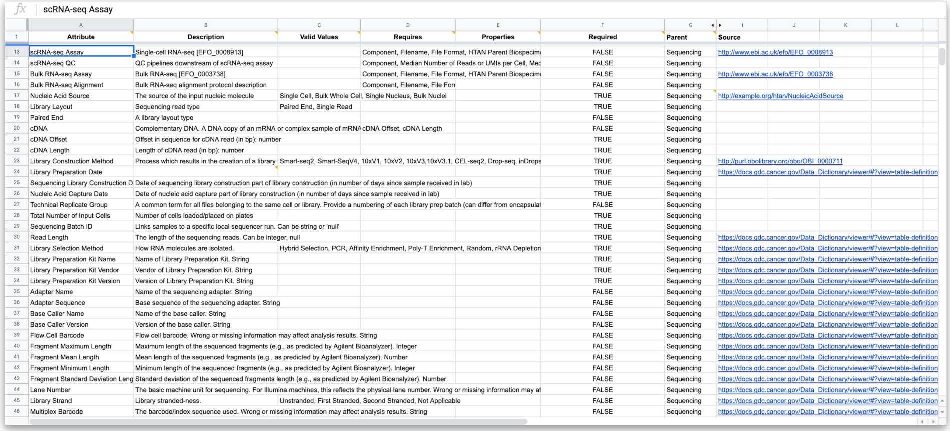
Last edit was made seconds ago by Milen Nikolov

Table 1. Proposed HTAN Metadata Model for sc/snRNA-Seq

Attribute	Description	Valid Values	Requires	Required	Suggested use cases	Source Ontology
HTAN Parent Biospecimen ID	HTAN Biospecimen Identifier from which these files were derived [https://docs.google.com/document/d/1podtPP8L1UNvVxx9_c_szlDcU1f8n7bige6XA_GoRVM/edit?usp=sharing]			TRUE	data integration	HTAN
HTAN ID	HTAN ID of this file HTANID SOP [https://docs.google.com/document/d/1podtPP8L1UNvVxx9_c_szlDcU1f8n7bige6XA_GoRVM/edit?usp=sharing]			TRUE	data integration with downstream levels	HTAN
Nucleic Acid Source	The source of the input nucleic molecule	Single Cell, Bulk Cell, Single Nucleus, Bulk Nuclei		TRUE	matching analysis tools; data discovery filter (e.g. scRNA-seq vs. snRNA-seq)	HTAN
Single Cell Dissociation Date	The date when the sample was dissociated and cells were isolated	DD/MM/YYYY		TRUE	Batch effect resolution	HTAN
Dissociation Method	The tissue dissociation method used for scRNASeq assays	gentleMACS		TRUE	Batch effect resolution	HTAN
Single Cell Isolation Method	The method by which cells are isolated into individual reaction containers at a single cell resolution (e.g. wells, micro-droplets)	Microfluidics Chip, Droplets, FACS		TRUE	batch effects resolution; matching analysis tools (e.g. enabling gene expression or splicing variants analysis)	HCA

Community driven specification
(RFC process in working groups)

Data curators translate into
model definition UI sheet

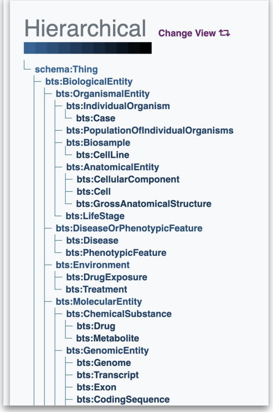


scRNA-seq Assay

	A	B	C	D	E	F	G	I	J	K	L
1	Attribute	Description	Valid Values	Requires	Properties	Required	Parent	Source			
13	scRNA-seq Assay	Single-cell RNA-seq [EFO_0008913]		Component, Filename, File Format, HTAN Parent Biospecim		FALSE	Sequencing	http://www.ebi.ac.uk/efo/EFO_0008913			
14	scRNA-seq QC	QC pipelines downstream of scRNA-seq assay		Component, Median Number of Reads or UMIs per Cell, Mec		FALSE	Sequencing				
15	Bulk RNA-seq Assay	Bulk RNA-seq [EFO_0003738]		Component, Filename, File Format, HTAN Parent Biospecim		FALSE	Sequencing	http://www.ebi.ac.uk/efo/EFO_0003738			
16	Bulk RNA-seq Alignment	Bulk RNA-seq alignment protocol description		Component, Filename, File Forr		FALSE	Sequencing				
17	Nucleic Acid Source	The source of the input nucleic molecule	Single Cell, Bulk Whole Cell, Single Nucleus, Bulk Nuclei			TRUE	Sequencing	http://example.org/htan/NucleicAcidSource			
18	Library Layout	Sequencing read type	Paired End, Single Read			TRUE	Sequencing				
19	Paired End	A library layout type				FALSE	Sequencing				
20	cDNA	Complementary DNA. A DNA copy of an mRNA or complex sample of mRNA		cDNA Offset, cDNA Length		FALSE	Sequencing				
21	cDNA Offset	Offset in sequence for cDNA read (in bp): number				TRUE	Sequencing				
22	cDNA Length	Length of cDNA read (in bp): number				TRUE	Sequencing				
23	Library Construction Method	Process which results in the creation of a library	Smart-seq2, Smart-SeqV4, 10xV1, 10xV2, 10xV3,10xV3.1, CEL-seq2, Drop-seq, inDrops			TRUE	Sequencing	http://purl.obolibrary.org/obo/OBI_0000711			
24	Library Preparation Date					TRUE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
25	Sequencing Library Construction D	Date of sequencing library construction part of library construction (in number of days since sample received in lab)				TRUE	Sequencing				
26	Nucleic Acid Capture Date	Date of nucleic acid capture part of library construction (in number of days since sample received in lab)				TRUE	Sequencing				
27	Technical Replicate Group	A common term for all files belonging to the same cell or library. Provide a numbering of each library prep batch (can differ from encapsulat				FALSE	Sequencing				
28	Total Number of Input Cells	Number of cells loaded/placed on plates				TRUE	Sequencing				
29	Sequencing Batch ID	Links samples to a specific local sequencer run. Can be string or 'null'				TRUE	Sequencing				
30	Read Length	The length of the sequencing reads. Can be integer, null				TRUE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
31	Library Selection Method	How RNA molecules are isolated.	Hybrid Selection, PCR, Affinity Enrichment, Poly-T Enrichment, Random, rRNA Depletion			TRUE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
32	Library Preparation Kit Name	Name of Library Preparation Kit. String				TRUE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
33	Library Preparation Kit Vendor	Vendor of Library Preparation Kit. String				TRUE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
34	Library Preparation Kit Version	Version of Library Preparation Kit. String				TRUE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
35	Adapter Name	Name of the sequencing adapter. String				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
36	Adapter Sequence	Base sequence of the sequencing adapter. String				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
37	Base Caller Name	Name of the base caller. String				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
38	Base Caller Version	Version of the base caller. String				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
39	Flow Cell Barcode	Flow cell barcode. Wrong or missing information may affect analysis results. String				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
40	Fragment Maximum Length	Maximum length of the sequenced fragments (e.g., as predicted by Agilent Bioanalyzer). Integer				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
41	Fragment Mean Length	Mean length of the sequenced fragments (e.g., as predicted by Agilent Bioanalyzer). Number				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
42	Fragment Minimum Length	Minimum length of the sequenced fragments (e.g., as predicted by Agilent Bioanalyzer). Integer				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
43	Fragment Standard Deviation Leng	Standard deviation of the sequenced fragments length (e.g., as predicted by Agilent Bioanalyzer). Number				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
44	Lane Number	The basic machine unit for sequencing. For Illumina machines, this reflects the physical lane number. Wrong or missing information may af				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
45	Library Strand	Library stranded-ness.	Unstranded, First Stranded, Second Stranded, Not Applicable			FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			
46	Multiplex Barcode	The barcode/index sequence used. Wrong or missing information may affect analysis results. String				FALSE	Sequencing	https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition			

Attribute, Description, **Valid Values**,
Requires, Properties, **Required**, Parent,
Source, **DependsOn**, **Validation Rules**

schematic
converts it to
JSON-LD schema



Hierarchical

- schema:Thing
  - bts:BiologicalEntity
    - bts:OrganismalEntity
      - bts:IndividualOrganism
        - bts:Case
      - bts:PopulationOfIndividualOrganisms
      - bts:Biosample
        - bts:CellLine
      - bts:AnatomicalEntity
        - bts:CellularComponent
        - bts:Cell
        - bts:GrossAnatomicalStructure
      - bts:LifeStage
    - bts:DiseaseOrPhenotypicFeature
      - bts:Disease
      - bts:PhenotypicFeature
    - bts:Environment
      - bts:DrugExposure
      - bts:Treatment
    - bts:MolecularEntity
      - bts:ChemicalSubstance
        - bts:Drug
        - bts:Metabolite
      - bts:GenomicEntity
        - bts:Genome
        - bts:Transcript
        - bts:Exon
        - bts:CodingSequence

Supporting many clinical, sample and assay types

Participant

- Demographics
- Diagnosis
- Family history
- Exposure
- Follow up
- Therapy
- Molecular test
- Cancer-specific
- Biospecimen

Single cell/nucleus sequencing

- sc/snRNA-seq
- sc/snATAC-seq
- sc/snMethyl-C-seq

Bulk sequencing

- Bulk DNA-seq
- Bulk RNA-seq
- Hi-C (in dev)

Mass Spectrometry

- Proteomics
- Metabolomics
- Lipidomics

Spatial Transcriptomics

- 10x Visium (in dev)
- Ex-Seq (in dev)
- MERFISH (in dev)

Multiplex Imaging

- CyCIF
- MxIF
- mIHC
- CODEX
- MIBI
- IMC
- SABER

Other Imaging

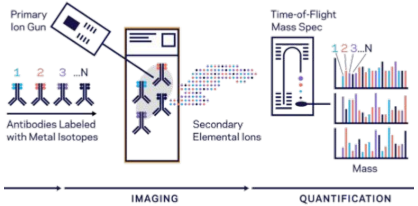
- H&E
- FIB-SEM (in dev)

Preparing (imaging) data for the DCC

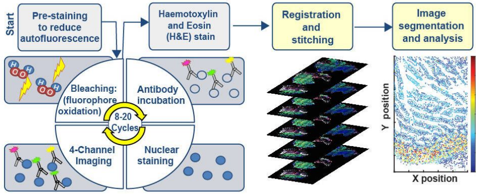
SageBionetworks

Explosion in imaging data size and diversity

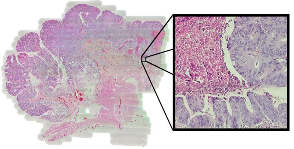
MIBI & IMC



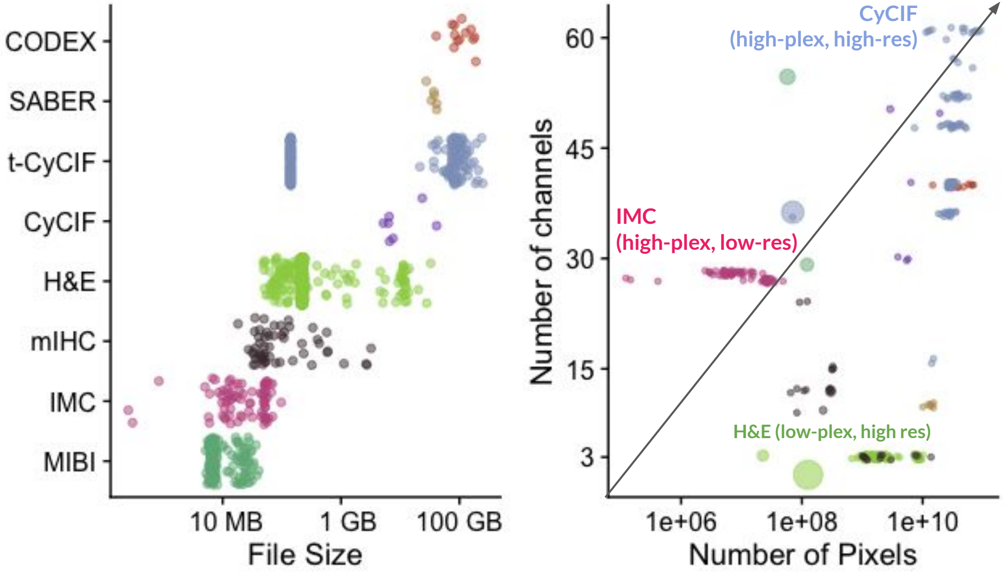
Cyclic (CyCIF, MxIF, mIHC)



Digital pathology (H&E, IHC)



And more: incl EM & clinical imaging



Increasingly larger, higher dimension and higher resolution data

Data types expected from Lung PCA / HTAN BU

(from the July 2021 data survey so may be out of date)

Additional assay type expected to submit	Timeframe for submission to the DCC	Assay will be included in manuscript submission? (Center: Please select Yes or No)	Assay covered by existing RFCs? (DCC to select. Center please comment in Column F if value looks wrong)	Center Contact (Center please provide a suggested contact for assay in highlighted cells)
Bulk RNA Sequencing (total RNA)	as early as Jan 2022	Yes		Sarah Mazzilli / Jennifer Beane
LCM- Whole Exome Sequencing	as early as Jan 2022	Yes		Sarah Mazzilli / Linh Tran
Multiplex immunofluorescence (mIF)	as early as Jan 2022	Yes		Sarah Mazzilli / Edwin Parra
H&E images	as early as Jan 2022	Yes		Sarah Mazzilli / Dan Merrick
Single cell RNA seq	as early as Jan 2022	Yes		Sarah Mazzilli / Josh Campbell
Single-cell DNA seq	possibly end of 2022- early 2023	Yes		Sarah Mazzilli/ Matthew Meyerson
CODEX	as early as Jan 2022	Yes		Sarah Mazzilli/ Wilson
Clinical & Biospecimen data	end of 2021	No		Sarah Mazzilli

Data released from Lung PCA / HTAN BU

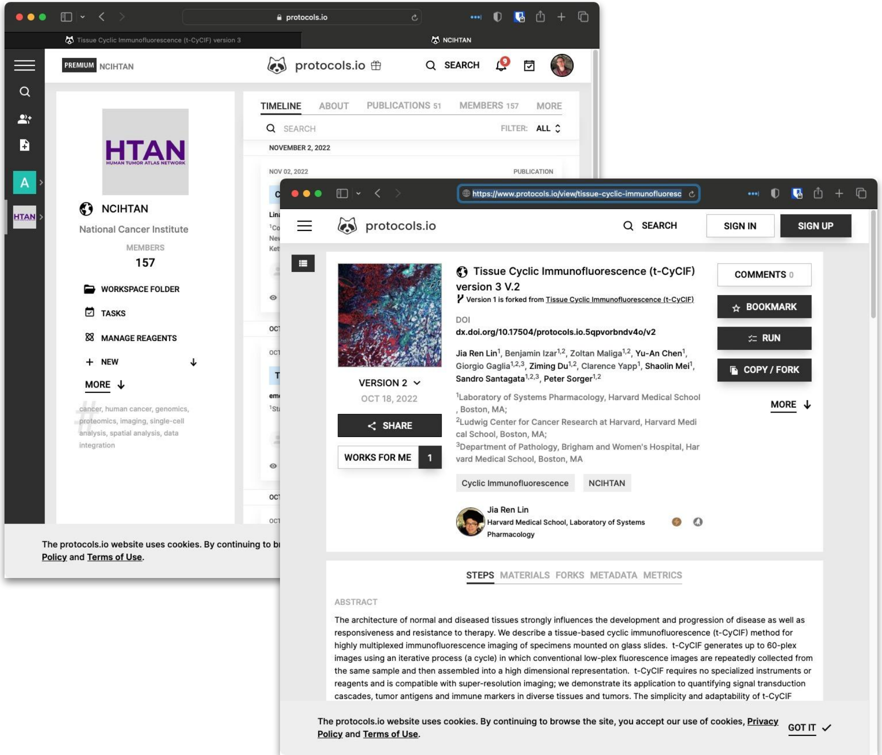
- 1 organ,
- 1 cancer type,
- 6 cases,
- 15 biospecimens,
- 1 assay,
- 177 files
  - 100 scRNA-seq Level 1,
  - ~~15 scRNA-seq Level 2~~ (currently unreleased - known DCC issue)
  - 45 scRNA-seq Level 3
  - 32 scRNA-seq Level 4

NCIHTAN workspace on protocols.io

protocols.io/workspaces/ncihtan

Request membership from DCC if required for upload/editing

protocols.io protocol or open-access publication required for many data templates



The protocols.io website uses cookies. By continuing to b
Policy and Terms of Use.

Tissue Cyclic Immunofluorescence (t-CyCIF) version 3 V.2

Version 1 is forked from Tissue Cyclic Immunofluorescence (t-CyCIF)

DOI
dx.doi.org/10.17504/protocols.io.5qpvorbndv4o/v2

Jia Ren Lin[1], Benjamin Izar[1,2], Zoltan Maliga[1,2], Yu-An Chen[1], Giorgio Gaglia[1,2,3], Ziming Du[1,2], Clarence Yapp[1], Shaolin Mei[1], Sandro Santagata[1,2,3], Peter Sorger[1,2]

[1]Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA;
[2]Ludwig Center for Cancer Research at Harvard, Harvard Medical School, Boston, MA;
[3]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

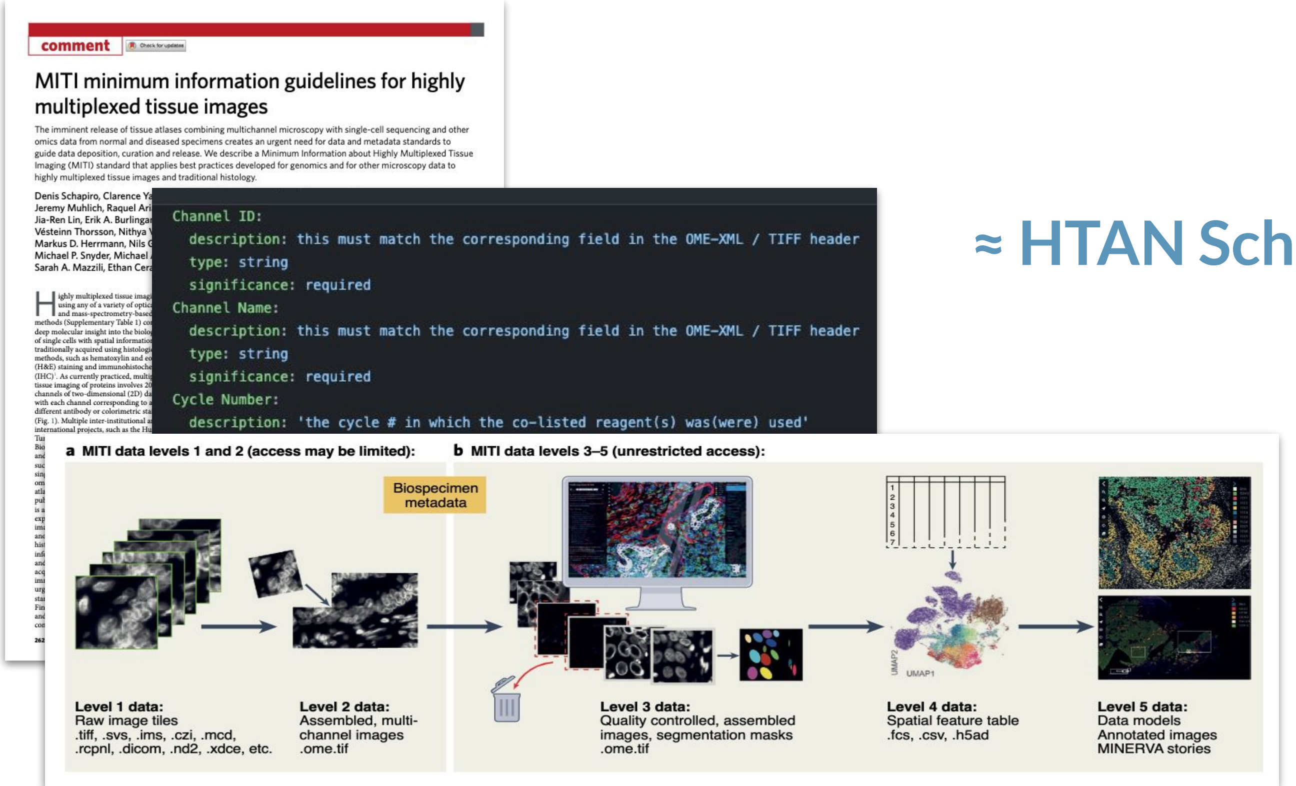
Jia Ren Lin
Harvard Medical School, Laboratory of Systems Pharmacology

ABSTRACT

The architecture of normal and diseased tissues strongly influences the development and progression of disease as well as responsiveness and resistance to therapy. We describe a tissue-based cyclic immunofluorescence (t-CyCIF) method for highly multiplexed immunofluorescence imaging of specimens mounted on glass slides. t-CyCIF generates up to 60-plex images using an iterative process (a cycle) in which conventional low-plex fluorescence images are repeatedly collected from the same sample and then assembled into a high dimensional representation. t-CyCIF requires no specialized instruments or reagents and is compatible with super-resolution imaging; we demonstrate its application to quantifying signal transduction cascades, tumor antigens and immune markers in diverse tissues and tumors. The simplicity and adaptability of t-CyCIF

The protocols.io website uses cookies. By continuing to browse the site, you accept our use of cookies, Privacy Policy and Terms of Use.

Minimum information for tissue imaging



comment

MITI minimum information guidelines for highly multiplexed tissue images

The imminent release of tissue atlases combining multichannel microscopy with single-cell sequencing and other omics data from normal and diseased specimens creates an urgent need for data and metadata standards to guide data deposition, curation and release. We describe a Minimum Information about Highly Multiplexed Tissue Imaging (MITI) standard that applies best practices developed for genomics and for other microscopy data to highly multiplexed tissue images and traditional histology.

Denis Schapiro, Clarence Ya
Jeremy Muhlich, Raquel Ari
Jia-Ren Lin, Erik A. Burlingar
Vésteinn Thorsson, Nithya V
Markus D. Herrmann, Nils G
Michael P. Snyder, Michael A
Sarah A. Mazzili, Ethan Cera

```
Channel ID:
  description: this must match the corresponding field in the OME-XML / TIFF header
  type: string
  significance: required
Channel Name:
  description: this must match the corresponding field in the OME-XML / TIFF header
  type: string
  significance: required
Cycle Number:
  description: 'the cycle # in which the co-listed reagent(s) was(were) used'
```

≈ **HTAN Schema**

HTAN Data Model Components imaging

Imaging Level 1

Raw imaging data, from the microscope, or intermediate processing steps. With supplementary non-imaging files as required

Stored as a compressed archive

77 imaging Attributes

39 required

24 with valid values

Imaging Level 2

TIFF based single file per image following any necessary pre-processing such as stitching, registration, illumination corraction.

Valid, tiled and pyramidal OME-TIFF strongly prefered

Unvalidated channel metadata is provided as a CSV file

Imaging Level 3 Segmentation

Mask, probability map or ROI point or polygon outlinng the location of features that will be summarised in Level 4.

Imaging Level 4

object x feature array containing summarty statistics per detected

Imaging Level 3 Image

Quality controlled image with required and validated OME-TIFF header elements

Points towards imaging level 3 channels for valid metadata

Imaging Level 3 Channels

Validated table of channel metadata capturing varaibles key to data integrety and filtering including antibody target and RRID identifiers

HTAN Data Model Components related to

Imaging Level 1

Component, Filename, File Format, HTAN Parent Biospecimen ID, HTAN Data File ID, Imaging Assay Type, Protocol Link, Software and Version, Commit SHA, Pre-processing Completed, Pre-processing Required, Comment

77 imaging Attributes

39 required

24 with valid values

Imaging Level 2

Component, Filename, File Format, HTAN Participant ID, HTAN Parent Biospecimen ID, HTAN Data File ID, Channel Metadata Filename, Imaging Assay Type, Protocol Link, Software and Version, Microscope, Objective, NominalMagnification, LensNA, WorkingDistance,WorkingDistanceUnit, Immersion, Pyramid, Zstack, Tseries, Passed QC, Comment, FOV number, FOVX, FOVXUnit, FOVY, FOVYUnit, Frame Averaging, Image ID, DimensionOrder, PhysicalSizeX, PhysicalSizeXUnit, PhysicalSizeY, PhysicalSizeYUnit, PhysicalSizeZ, PhysicalSizeZUnit, Pixels BigEndian, PlaneCount, SizeC, SizeT, SizeX, SizeY, SizeZ, PixelType

Imaging Level 3 Segmentation

Component, Filename, File Format, HTAN Parent Data File ID, HTAN Data File ID, Imaging Segmentation Data Type, Parameter file, Software and Version, Commit SHA, Imaging Object Class, Imaging Object Class Description, Number of Objects

Imaging Level 4

Component, Filename, File Format, HTAN Parent Data File ID, HTAN Parent Channel Metadata ID, HTAN Data File ID, Parameter file, Software and Version, Commit SHA,Number of Objects, Number of Features,Imaging Object Class, Imaging Object Class Description, Imaging Summary Statistic

Imaging Level 3 Image

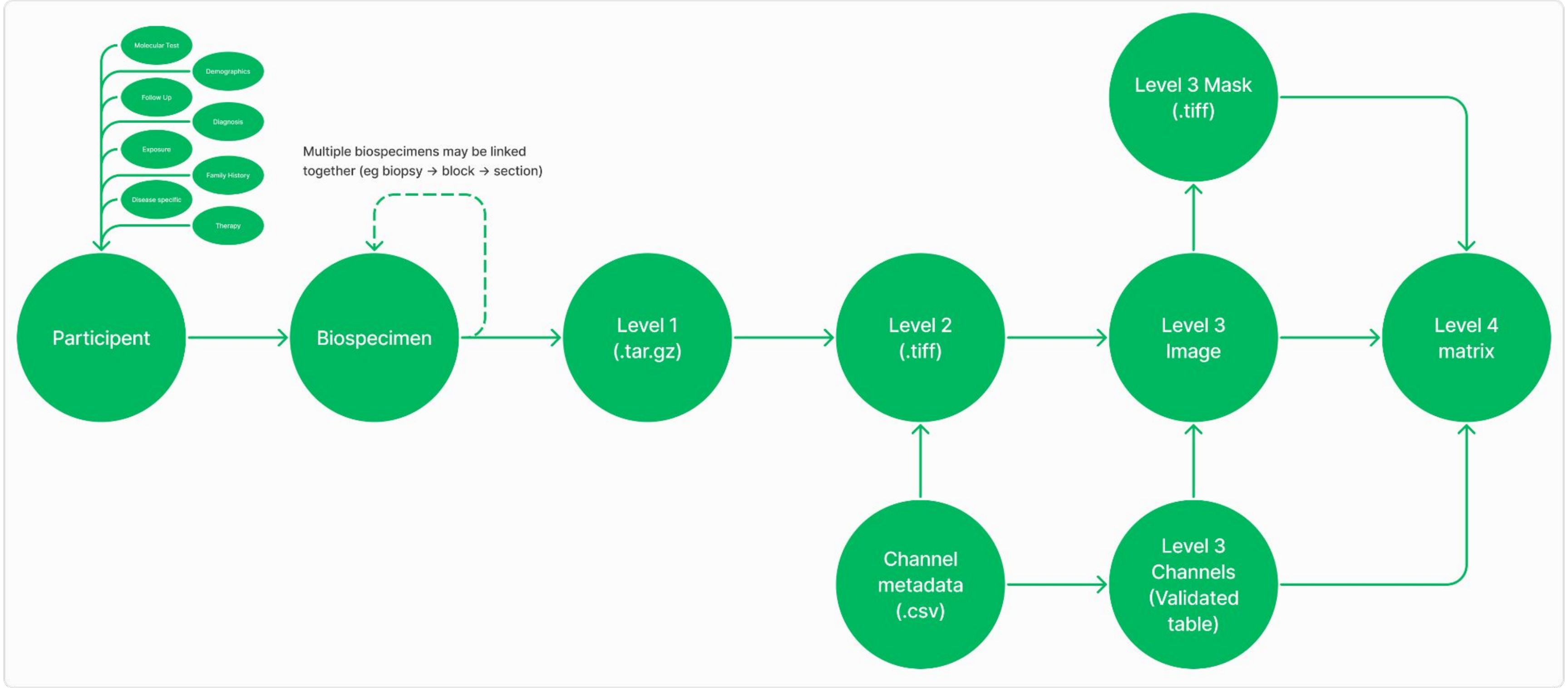
Component, Filename, File Format, HTAN Parent Biospecimen ID, HTAN Parent Data File ID, HTAN Parent Channel Metadata ID, HTAN Data File ID, Imaging Assay Type, Protocol Link,Software and Version, Microscope, Objective, NominalMagnification, LensNA, WorkingDistance, Immersion, Pyramid, Zstack, Tseries, Passed QC, Comment, FOV number, FOVX, FOVY, Frame Averaging

Imaging Level 3 Channels

HTAN Channel Metadata ID, Channel ID, Channel Name, Channel Passed QC, Cycle Number, Sub Cycle Number, Antibody Role, Target Name, Antibody Name, RRID identifier, Fluorophore, Clone, Lot, Vendor, Catalog Number, Excitation Wavelength, Emission Wavelength, Excitation Bandwidth, Emission Bandwidth, Metal Isotope Element, Metal Isotope Mass, Oligo Barcode Upper Strand, Oligo Barcode Lower Strand, Dilution, Concentration

Complete imaging dataset

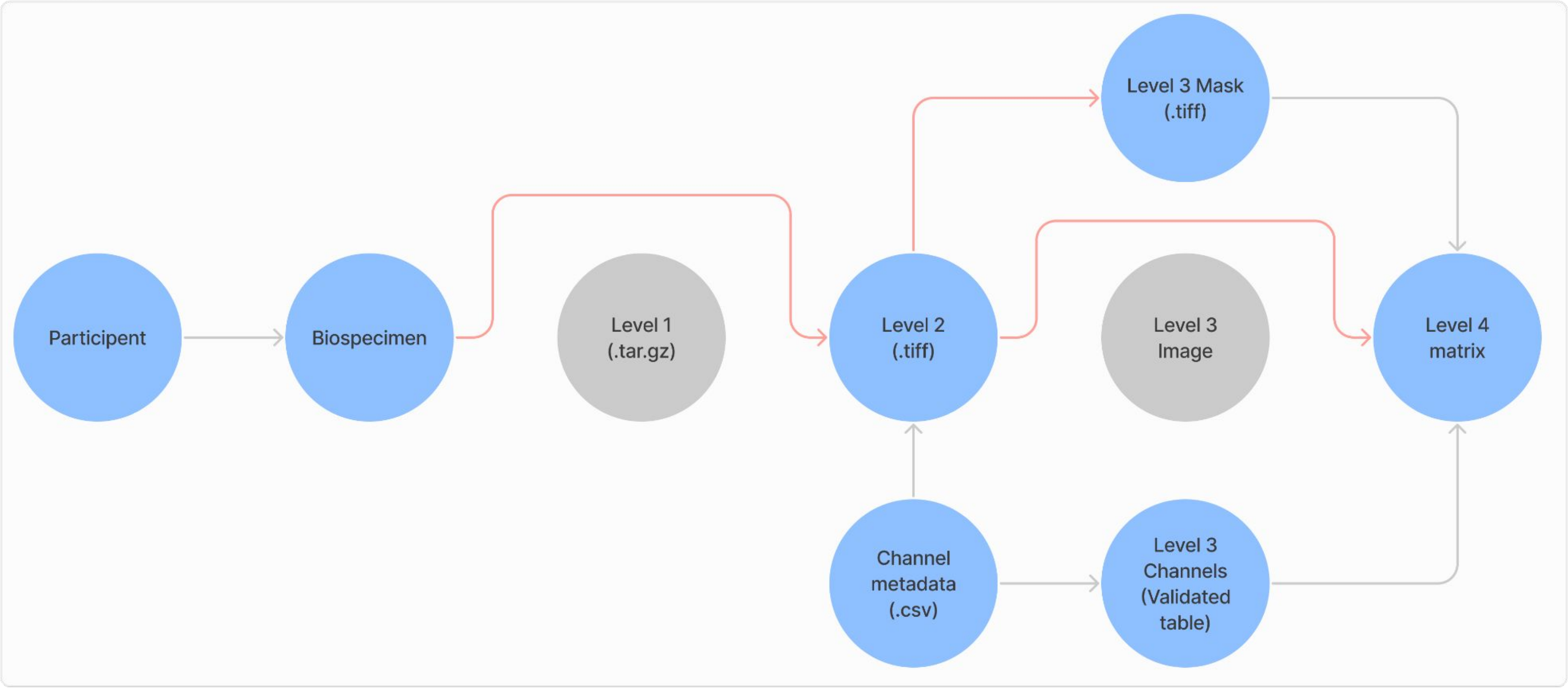
Ideal



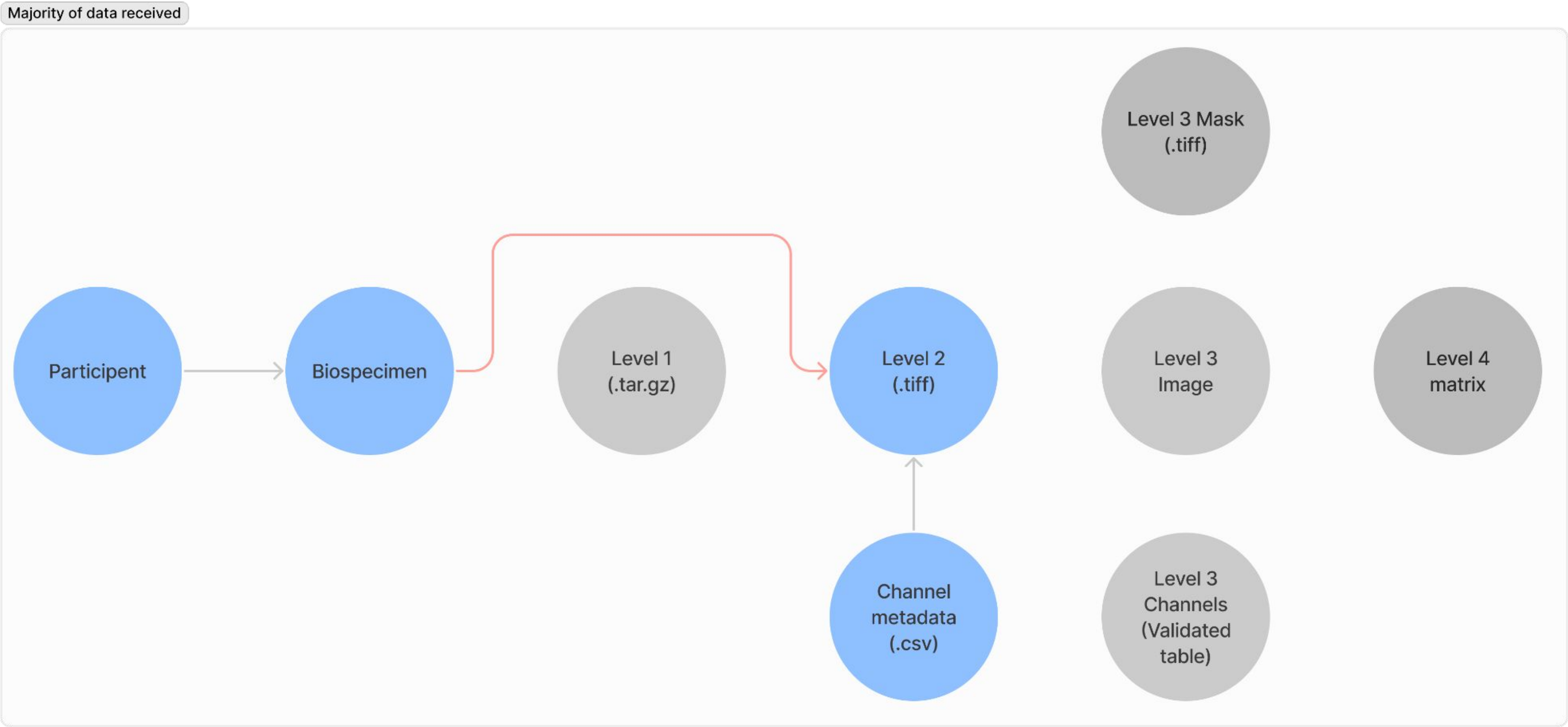
Multiple biospecimens may be linked together (eg biopsy → block → section)

A minimal imaging dataset

Lightweight



Majority of data received to date



File formats

We strongly prefer OME-TIFF images that are

- ✅ Valid OME-TIFFs with valid OME-XML
- 🔺 Contain a single image pyramid
- 🧩 Are tiled
- □ Contain informative channel names that match channel metadata file
    - e.g. ‘**Ki-67**’ rather than ‘**Channel15**’
- Do not contain acquisition dates, un-redacted slide labels or other potential PII/PHI

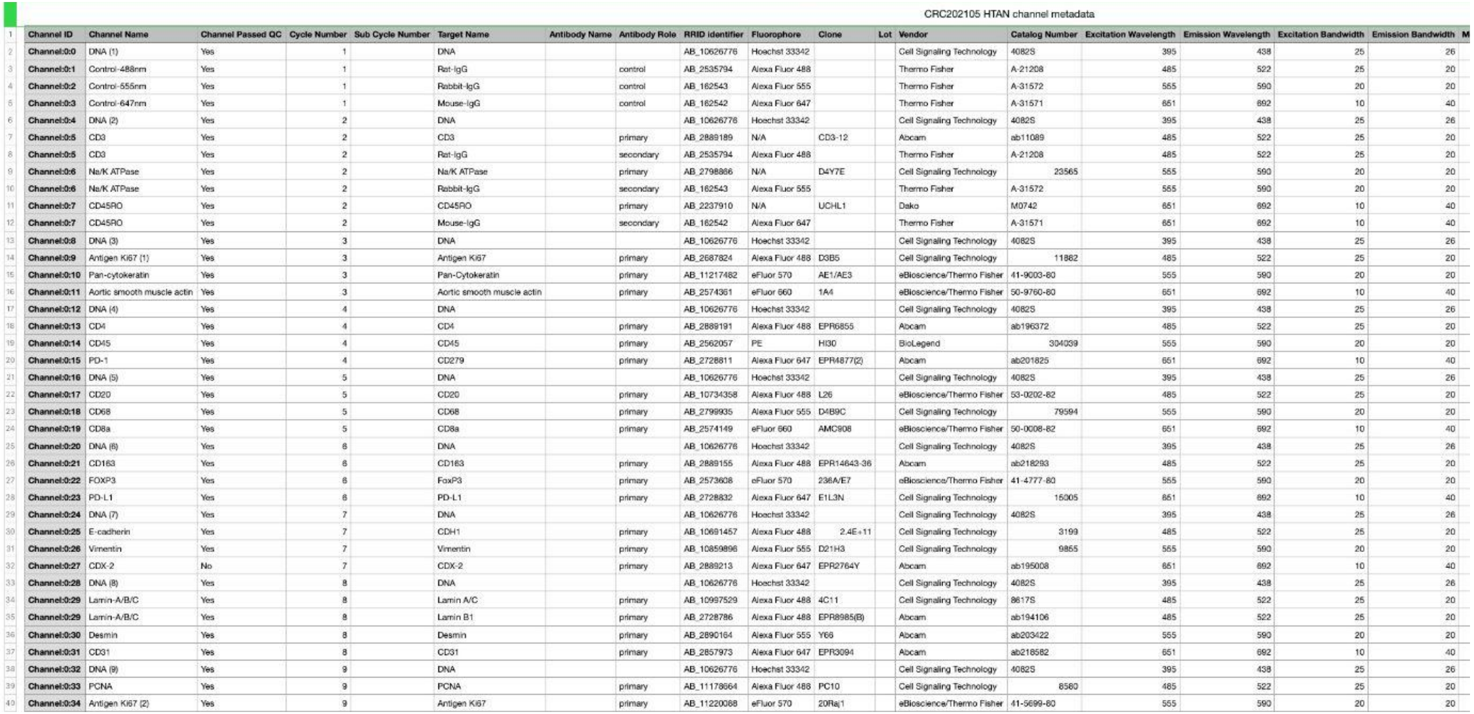
If OME-TIFF conversion is not possible it is possible to submit another Bioformats compatible image format. Please reach out to the DCC.

Channel metadata file

- Contains important information on channel metadata
    - Channel name
    - Channel QC
    - Antibody RRID
- Channel name should match file where possible
- Individual rows for primary and secondary antibodies where used
- Generate template in DCA using “Imaging Level 3 Channels”
- For imaging level 2 provide the folder path in Synapse to a csv

Channel metadata file example

syn25763824 | CRC202105 HTAN channel metadata.csv | HTAN HMS



CRC202105 HTAN channel metadata

Channel ID	Channel Name	Channel Passed QC	Cycle Number	Sub Cycle Number	Target Name	Antibody Name	Antibody Role	RRID identifier	Fluorophore	Clone	Lot	Vendor	Catalog Number	Excitation Wavelength	Emission Wavelength	Excitation Bandwidth	Emission Bandwidth	M
Channel:0:0	DNA (1)	Yes	1		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:1	Control-488nm	Yes	1		Rat-IgG		control	AB_2535794	Alexa Fluor 488			Thermo Fisher	A-21208	485	522	25	20	
Channel:0:2	Control-555nm	Yes	1		Rabbit-IgG		control	AB_162543	Alexa Fluor 555			Thermo Fisher	A-31572	555	590	20	20	
Channel:0:3	Control-647nm	Yes	1		Mouse-IgG		control	AB_162542	Alexa Fluor 647			Thermo Fisher	A-31571	651	692	10	40	
Channel:0:4	DNA (2)	Yes	2		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:5	CD3	Yes	2		CD3		primary	AB_2889189	N/A	CD3-12		Abcam	ab11089	485	522	25	20	
Channel:0:5	CD3	Yes	2		Rat-IgG		secondary	AB_2535794	Alexa Fluor 488			Thermo Fisher	A-21208	485	522	25	20	
Channel:0:6	Na/K ATPase	Yes	2		Na/K ATPase		primary	AB_2798886	N/A	D4Y7E		Cell Signaling Technology	23565	555	590	20	20	
Channel:0:6	Na/K ATPase	Yes	2		Rabbit-IgG		secondary	AB_162543	Alexa Fluor 555			Thermo Fisher	A-31572	555	590	20	20	
Channel:0:7	CD45RO	Yes	2		CD45RO		primary	AB_2237910	N/A	UCHL1		Dako	M0742	651	692	10	40	
Channel:0:7	CD45RO	Yes	2		Mouse-IgG		secondary	AB_162542	Alexa Fluor 647			Thermo Fisher	A-31571	651	692	10	40	
Channel:0:8	DNA (3)	Yes	3		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:9	Antigen Ki67 (1)	Yes	3		Antigen Ki67		primary	AB_2687824	Alexa Fluor 488	D3B5		Cell Signaling Technology	11882	485	522	25	20	
Channel:0:10	Pan-cytokeratin	Yes	3		Pan-Cytokeratin		primary	AB_11217482	eFluor 570	AE1/AE3		eBioscience/Thermo Fisher	41-9003-80	555	590	20	20	
Channel:0:11	Aortic smooth muscle actin	Yes	3		Aortic smooth muscle actin		primary	AB_2574361	eFluor 660	1A4		eBioscience/Thermo Fisher	50-9760-80	651	692	10	40	
Channel:0:12	DNA (4)	Yes	4		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:13	CD4	Yes	4		CD4		primary	AB_2889191	Alexa Fluor 488	EPR6855		Abcam	ab196372	485	522	25	20	
Channel:0:14	CD45	Yes	4		CD45		primary	AB_2562057	PE	HI30		BioLegend	304039	555	590	20	20	
Channel:0:15	PD-1	Yes	4		CD279		primary	AB_2728811	Alexa Fluor 647	EPR4877(2)		Abcam	ab201825	651	692	10	40	
Channel:0:16	DNA (5)	Yes	5		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:17	CD20	Yes	5		CD20		primary	AB_10734358	Alexa Fluor 488	L26		eBioscience/Thermo Fisher	53-0202-82	485	522	25	20	
Channel:0:18	CD68	Yes	5		CD68		primary	AB_2799935	Alexa Fluor 555	D4B9C		Cell Signaling Technology	79594	555	590	20	20	
Channel:0:19	CD8a	Yes	5		CD8a		primary	AB_2574149	eFluor 660	AMC908		eBioscience/Thermo Fisher	50-0008-82	651	692	10	40	
Channel:0:20	DNA (6)	Yes	6		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:21	CD163	Yes	6		CD163		primary	AB_2889155	Alexa Fluor 488	EPR14643-36		Abcam	ab218293	485	522	25	20	
Channel:0:22	FOXP3	Yes	6		FoxP3		primary	AB_2573608	eFluor 570	236A/E7		eBioscience/Thermo Fisher	41-4777-80	555	590	20	20	
Channel:0:23	PD-L1	Yes	6		PD-L1		primary	AB_2728832	Alexa Fluor 647	E1L3N		Cell Signaling Technology	15005	651	692	10	40	
Channel:0:24	DNA (7)	Yes	7		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:25	E-cadherin	Yes	7		CDH1		primary	AB_10691457	Alexa Fluor 488	2.4E+11		Cell Signaling Technology	3199	485	522	25	20	
Channel:0:26	Vimentin	Yes	7		Vimentin		primary	AB_10859896	Alexa Fluor 555	D21H3		Cell Signaling Technology	9855	555	590	20	20	
Channel:0:27	CDX-2	No	7		CDX-2		primary	AB_2889213	Alexa Fluor 647	EPR2764Y		Abcam	ab195008	651	692	10	40	
Channel:0:28	DNA (8)	Yes	8		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:29	Lamin-A/B/C	Yes	8		Lamin A/C		primary	AB_10997529	Alexa Fluor 488	4C11		Cell Signaling Technology	8617S	485	522	25	20	
Channel:0:29	Lamin-A/B/C	Yes	8		Lamin B1		primary	AB_2728786	Alexa Fluor 488	EPR8985(B)		Abcam	ab194106	485	522	25	20	
Channel:0:30	Desmin	Yes	8		Desmin		primary	AB_2890164	Alexa Fluor 555	Y66		Abcam	ab203422	555	590	20	20	
Channel:0:31	CD31	Yes	8		CD31		primary	AB_2857973	Alexa Fluor 647	EPR3094		Abcam	ab218582	651	692	10	40	
Channel:0:32	DNA (9)	Yes	9		DNA			AB_10626776	Hoechst 33342			Cell Signaling Technology	4082S	395	438	25	26	
Channel:0:33	PCNA	Yes	9		PCNA		primary	AB_11178664	Alexa Fluor 488	PC10		Cell Signaling Technology	8580	485	522	25	20	
Channel:0:34	Antigen Ki67 (2)	Yes	9		Antigen Ki67		primary	AB_11220088	eFluor 570	20Raj1		eBioscience/Thermo Fisher	41-5699-80	555	590	20	20	

De-identification

- Data submitted to the DCC must be de-identified by the Safe Harbor method
- All participant identifiers must be removed from metadata **and data**
- Make sure your De-identification risk assessment is up to date and that you are familiar with it.
- Key identified risks
  - Slide labels: Misplaced slide label stickers and visible dates
  - Acquisition dates: Potential re-identification attack vector through other metadata such as ‘days to sectioning from index`
  - Other identifiers (eg patient numbers) in original file metadata

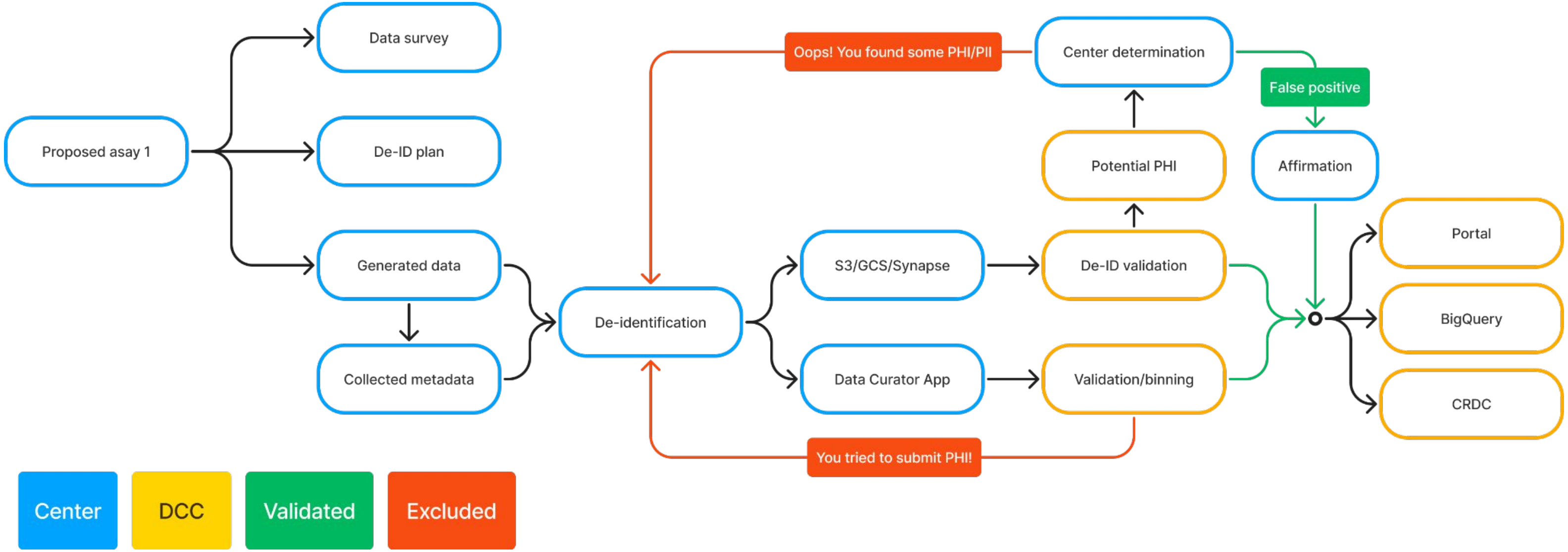
De-identification: Dates in imaging data

- **All dates must be removed, redacted or modified.**
- These include
    - DateTime tag,
    - AcquisitionDate in OME-XML
    - Date in Aperio ImageDescription
    - slide label images
    - or any other location
- Please check for and remove before submission to the DCC
  Even if they are not related to date of acquisition!
- DCC is validating De-identification and will withhold release

Reach out to me if you have issues identifying or removing potential PII/PHI in your imaging data

I can recommend tools and approaches

De-identification workflow

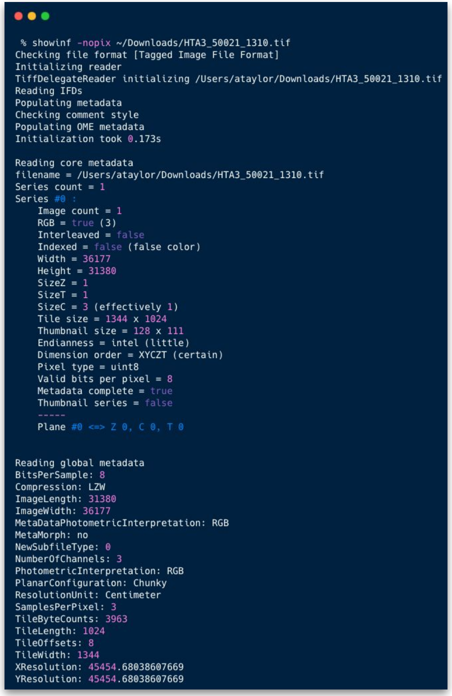


Gotchas

- Slide label sticker misplaced with participant DoB visible
- Date of acquisition included in OME-XML structured annotation original metadata
- 27 channel image encoded as 27 timepoints
- 40 channel CODEX encoded as 4 channels and 10 timepoints

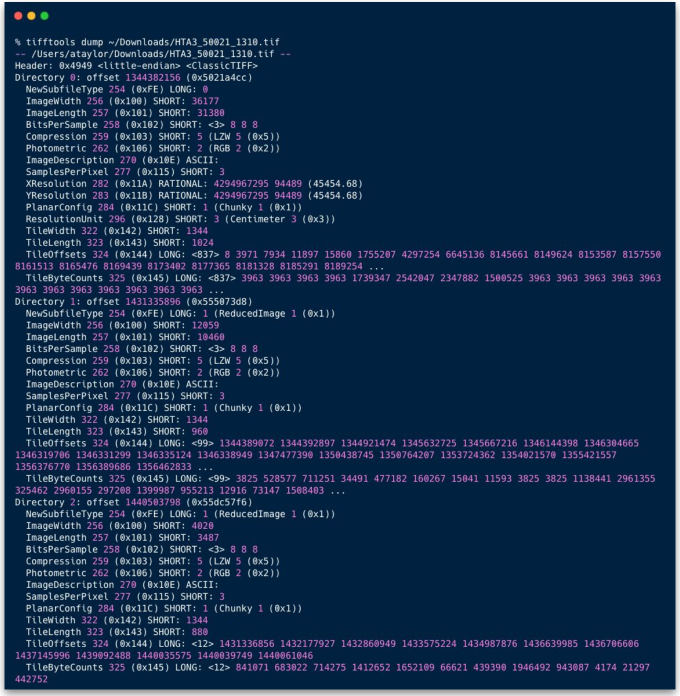
Useful checks before upload

showinf -nopix (bftools)



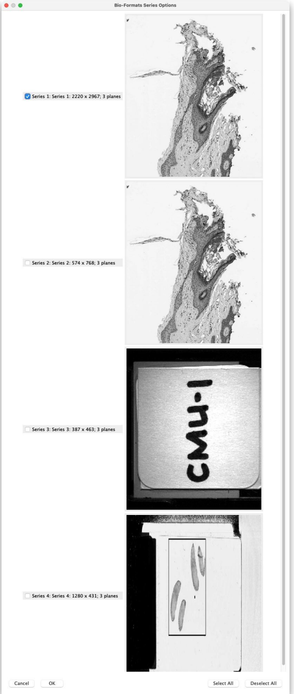
```
 % showinf -nopix ~/Downloads/HTA3_50021_1310.tif
Checking file format [Tagged Image File Format]
Initializing reader
TiffDelegateReader initializing /Users/ataylor/Downloads/HTA3_50021_1310.tif
Reading IFDs
Populating metadata
Checking comment style
Populating OME metadata
Initialization took 0.173s

Reading core metadata
filename = /Users/ataylor/Downloads/HTA3_50021_1310.tif
Series count = 1
Series #0 :
    Image count = 1
    RGB = true (3)
    Interleaved = false
    Indexed = false (false color)
    Width = 36177
    Height = 31380
    SizeZ = 1
    SizeT = 1
    SizeC = 3 (effectively 1)
    Tile size = 1344 x 1024
    Thumbnail size = 128 x 111
    Endianness = intel (little)
    Dimension order = XYCZT (certain)
    Pixel type = uint8
    Valid bits per pixel = 8
    Metadata complete = true
    Thumbnail series = false
    -----
    Plane #0 <=> Z 0, C 0, T 0


Reading global metadata
BitsPerSample: 8
Compression: LZW
ImageLength: 31380
ImageWidth: 36177
MetaDataPhotometricInterpretation: RGB
MetaMorph: no
NewSubfileType: 0
NumberOfChannels: 3
PhotometricInterpretation: RGB
PlanarConfiguration: Chunky
ResolutionUnit: Centimeter
SamplesPerPixel: 3
TileByteCounts: 3963
TileLength: 1024
TileOffsets: 8
TileWidth: 1344
XResolution: 45454.68038607669
YResolution: 45454.68038607669
```

tifftools dump



```
% tifftools dump ~/Downloads/HTA3_50021_1310.tif
-- /Users/ataylor/Downloads/HTA3_50021_1310.tif --
Header: 0x4949 <little-endian> <ClassicTIFF>
Directory 0: offset 1344382156 (0x5021a4cc)
  NewSubfileType 254 (0xFE) LONG: 0
  ImageWidth 256 (0x100) SHORT: 36177
  ImageLength 257 (0x101) SHORT: 31380
  BitsPerSample 258 (0x102) SHORT: <3> 8 8 8
  Compression 259 (0x103) SHORT: 5 (LZW 5 (0x5))
  Photometric 262 (0x106) SHORT: 2 (RGB 2 (0x2))
  ImageDescription 270 (0x10E) ASCII:
  SamplesPerPixel 277 (0x115) SHORT: 3
  XResolution 282 (0x11A) RATIONAL: 4294967295 94489 (45454.68)
  YResolution 283 (0x11B) RATIONAL: 4294967295 94489 (45454.68)
  PlanarConfig 284 (0x11C) SHORT: 1 (Chunky 1 (0x1))
  ResolutionUnit 296 (0x128) SHORT: 3 (Centimeter 3 (0x3))
  TileWidth 322 (0x142) SHORT: 1344
  TileLength 323 (0x143) SHORT: 1024
  TileOffsets 324 (0x144) LONG: <837> 8 3971 7934 11897 15860 1755207 4297254 6645136 8145661 8149624 8153587 8157550
8161513 8165476 8169439 8173402 8177365 8181328 8185291 8189254 ...
  TileByteCounts 325 (0x145) LONG: <837> 3963 3963 3963 3963 1739347 2542047 2347882 1500525 3963 3963 3963 3963 3963
3963 3963 3963 3963 3963 3963 3963 ...
Directory 1: offset 1431335896 (0x555073d8)
  NewSubfileType 254 (0xFE) LONG: 1 (ReducedImage 1 (0x1))
  ImageWidth 256 (0x100) SHORT: 12059
  ImageLength 257 (0x101) SHORT: 10460
  BitsPerSample 258 (0x102) SHORT: <3> 8 8 8
  Compression 259 (0x103) SHORT: 5 (LZW 5 (0x5))
  Photometric 262 (0x106) SHORT: 2 (RGB 2 (0x2))
  ImageDescription 270 (0x10E) ASCII:
  SamplesPerPixel 277 (0x115) SHORT: 3
  PlanarConfig 284 (0x11C) SHORT: 1 (Chunky 1 (0x1))
  TileWidth 322 (0x142) SHORT: 1344
  TileLength 323 (0x143) SHORT: 960
  TileOffsets 324 (0x144) LONG: <99> 1344389072 1344392897 1344921474 1345632725 1345667216 1346144398 1346304665
1346319706 1346331299 1346335124 1346338949 1347477390 1350438745 1350764207 1353724362 1354021570 1355421557
1356376770 1356389686 1356462833 ...
  TileByteCounts 325 (0x145) LONG: <99> 3825 528577 711251 34491 477182 160267 15041 11593 3825 3825 1138441 2961355
325462 2960155 297208 1399987 955213 12916 73147 1508403 ...
Directory 2: offset 1440503798 (0x55dc57f6)
  NewSubfileType 254 (0xFE) LONG: 1 (ReducedImage 1 (0x1))
  ImageWidth 256 (0x100) SHORT: 4020
  ImageLength 257 (0x101) SHORT: 3487
  BitsPerSample 258 (0x102) SHORT: <3> 8 8 8
  Compression 259 (0x103) SHORT: 5 (LZW 5 (0x5))
  Photometric 262 (0x106) SHORT: 2 (RGB 2 (0x2))
  ImageDescription 270 (0x10E) ASCII:
  SamplesPerPixel 277 (0x115) SHORT: 3
  PlanarConfig 284 (0x11C) SHORT: 1 (Chunky 1 (0x1))
  TileWidth 322 (0x142) SHORT: 1344
  TileLength 323 (0x143) SHORT: 880
  TileOffsets 324 (0x144) LONG: <12> 1431336856 1432177927 1432860949 1433575224 1434987876 1436639985 1436706606
1437145996 1439092488 1440035575 1440039749 1440061046
  TileByteCounts 325 (0x145) LONG: <12> 841071 683022 714275 1412652 1652109 66621 439390 1946492 943087 4174 21297
442752
```

Bio-formats importer
(FIJI / ImageJ)



☑ Series 1: Series 1: 2220 x 2967; 3 planes

☐ Series 2: Series 2: 574 x 768; 3 planes

☐ Series 3: Series 3: 387 x 463; 3 planes

☐ Series 4: Series 4: 1280 x 431; 3 planes

CC-BY-4.0 licensing of all imaging data

- Imaging Data Commons enables free download of data from their cloud buckets thanks to the Google Public Datasets Program
- Requires that data are open access (otherwise download is requester-pays)
- DCC will request your approval for CC-BY-4.0 license to be applied to your data Before release on the portal and transfer to CDS
- This is in alignment with the HTAN External Data Sharing Policy
- Please reach out to DCC and Sean Hanlon (NCI PO) with any questions or concerns

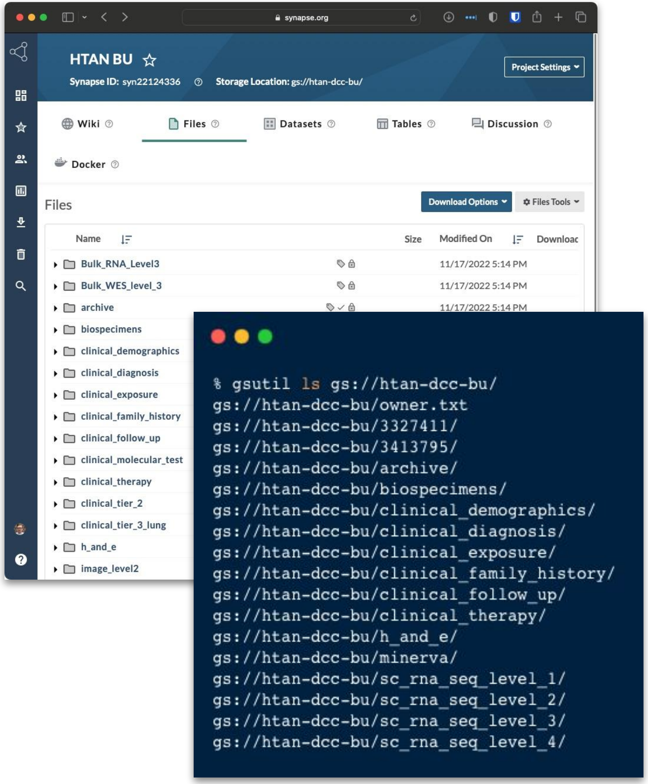
Submitting data and metadata to the HTAN DCC

SageBionetworks

Data organisation and upload

- Create top-level folders on Synapse
- DCC recommends that you have one top-level folder per data type and upload
- One manifest per top-level folder
- Data upload directly in Synapse or to bucket (BU is in Google Cloud Buckets)
- If key matches Synapse folder and md5 provided will by sync'd with Synapse

```
BASE64=$(openssl md5 -binary $file | base64)
gsutil -h x-goog-meta-content-md5:BASE64 \
     cp $file gs://htan-dcc-bu/imaging_level2/
```



HTAN BU

Synapse ID: syn22124336   Storage Location: gs://htan-dcc-bu/

Files

Name	Size	Modified On	Download
Bulk_RNA_Level3		11/17/2022 5:14 PM	
Bulk_WES_level_3		11/17/2022 5:14 PM	
archive		11/17/2022 5:14 PM	
biospecimens			
clinical_demographics			
clinical_diagnosis			
clinical_exposure			
clinical_family_history			
clinical_follow_up			
clinical_molecular_test			
clinical_therapy			
clinical_tier_2			
clinical_tier_3_lung			
h_and_e			
image_level2			

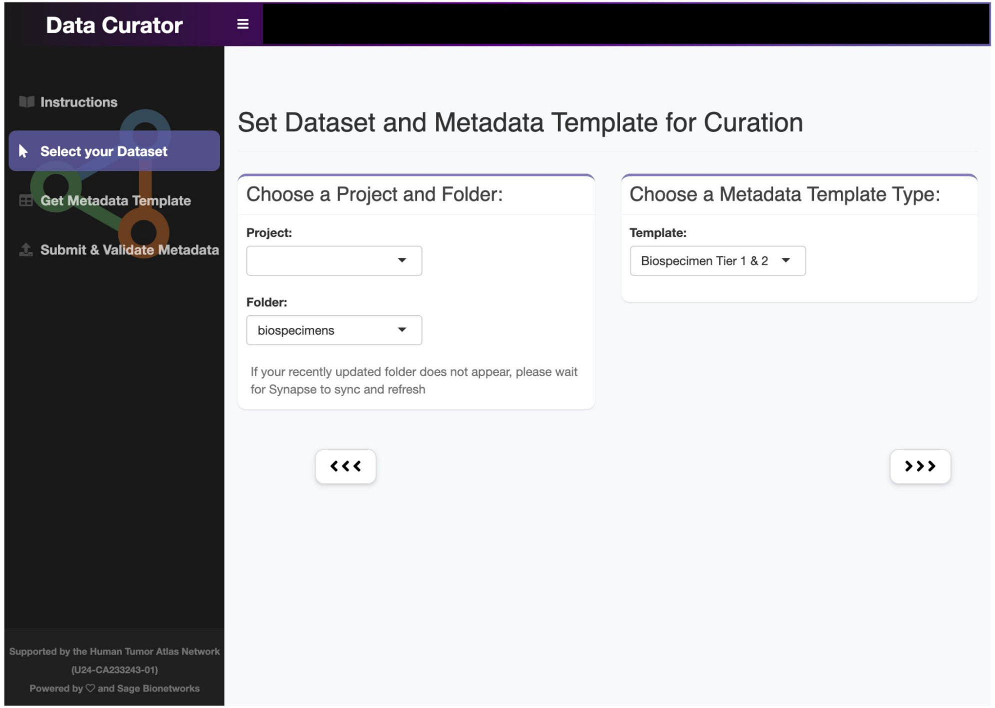
```
% gsutil ls gs://htan-dcc-bu/
gs://htan-dcc-bu/owner.txt
gs://htan-dcc-bu/3327411/
gs://htan-dcc-bu/3413795/
gs://htan-dcc-bu/archive/
gs://htan-dcc-bu/biospecimens/
gs://htan-dcc-bu/clinical_demographics/
gs://htan-dcc-bu/clinical_diagnosis/
gs://htan-dcc-bu/clinical_exposure/
gs://htan-dcc-bu/clinical_family_history/
gs://htan-dcc-bu/clinical_follow_up/
gs://htan-dcc-bu/clinical_therapy/
gs://htan-dcc-bu/h_and_e/
gs://htan-dcc-bu/minerva/
gs://htan-dcc-bu/sc_rna_seq_level_1/
gs://htan-dcc-bu/sc_rna_seq_level_2/
gs://htan-dcc-bu/sc_rna_seq_level_3/
gs://htan-dcc-bu/sc_rna_seq_level_4/
```

Let’s take a look at the Data Curator App

Data-model driven

Structured data - follows template sheets

Validates
- columns headers
- controlled vocabulary
- ID cross-check match



Set Dataset and Metadata Template for Curation

Choose a Project and Folder:

Project:

Folder: biospecimens

If your recently updated folder does not appear, please wait for Synapse to sync and refresh

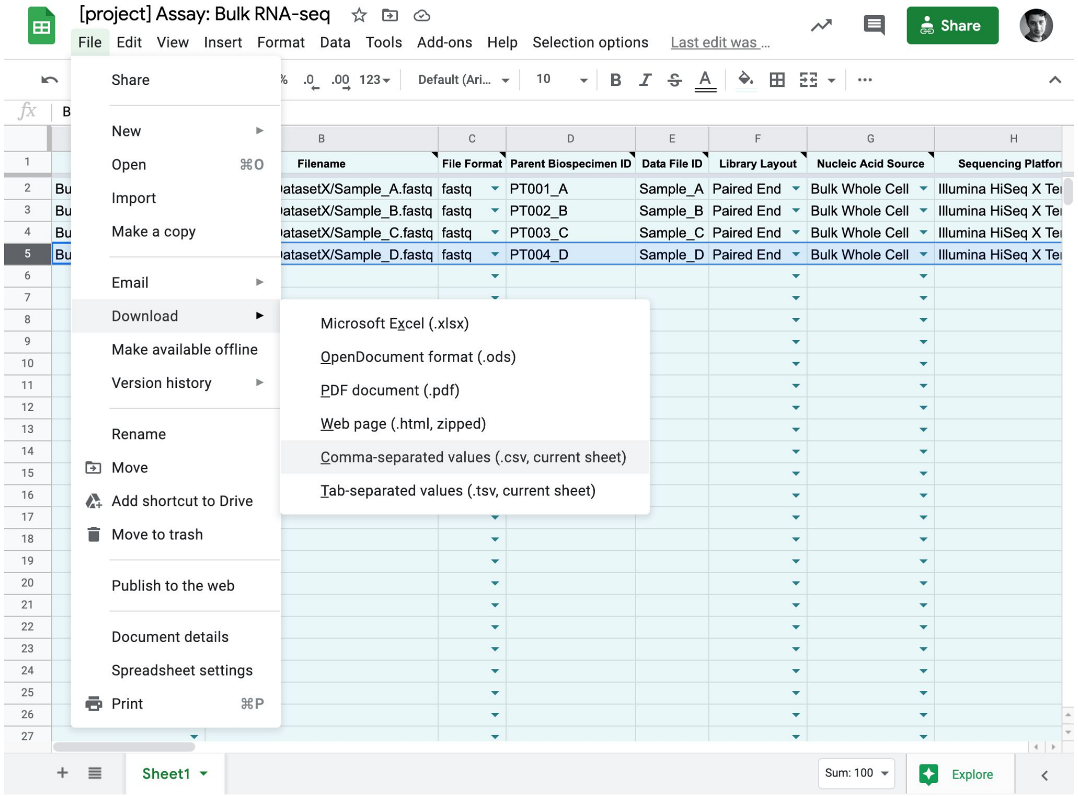
Choose a Metadata Template Type:

Template: Biospecimen Tier 1 & 2

Supported by the Human Tumor Atlas Network (U24-CA233243-01)
Powered by ♡ and Sage Bionetworks

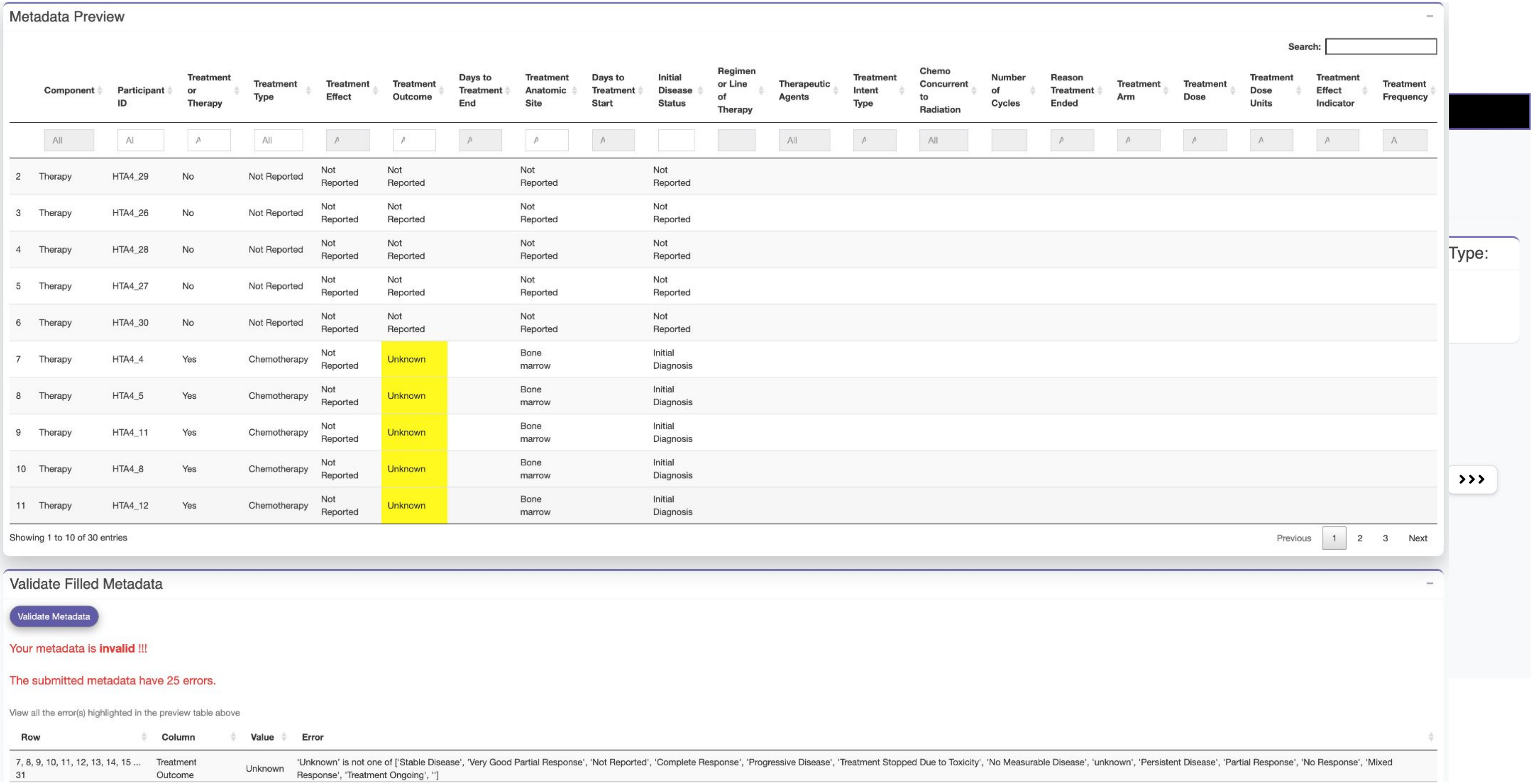
Interfaces w/ data model to generate templates

Can fill in online or locally in Excel/csv



	Filename	File Format	Parent Biospecimen ID	Data File ID	Library Layout	Nucleic Acid Source	Sequencing Platfor
Bu	atasetX/Sample_A.fastq	fastq	PT001_A	Sample_A	Paired End	Bulk Whole Cell	Illumina HiSeq X Te
Bu	atasetX/Sample_B.fastq	fastq	PT002_B	Sample_B	Paired End	Bulk Whole Cell	Illumina HiSeq X Te
Bu	atasetX/Sample_C.fastq	fastq	PT003_C	Sample_C	Paired End	Bulk Whole Cell	Illumina HiSeq X Te
Bu	atasetX/Sample_D.fastq	fastq	PT004_D	Sample_D	Paired End	Bulk Whole Cell	Illumina HiSeq X Te

Validates data for data-model compliance



Metadata Preview

Search:

	Component	Participant ID	Treatment or Therapy	Treatment Type	Treatment Effect	Treatment Outcome	Days to Treatment End	Treatment Anatomic Site	Days to Treatment Start	Initial Disease Status	Regimen or Line of Therapy	Therapeutic Agents	Treatment Intent Type	Chemo Concurrent to Radiation	Number of Cycles	Reason Treatment Ended	Treatment Arm	Treatment Dose	Treatment Dose Units	Treatment Effect Indicator	Treatment Frequency
	All	Al	A	All	A	A	A	A	A			All	A	All		A	A	A	A	A	A
2	Therapy	HTA4_29	No	Not Reported	Not Reported	Not Reported		Not Reported		Not Reported											
3	Therapy	HTA4_26	No	Not Reported	Not Reported	Not Reported		Not Reported		Not Reported											
4	Therapy	HTA4_28	No	Not Reported	Not Reported	Not Reported		Not Reported		Not Reported											
5	Therapy	HTA4_27	No	Not Reported	Not Reported	Not Reported		Not Reported		Not Reported											
6	Therapy	HTA4_30	No	Not Reported	Not Reported	Not Reported		Not Reported		Not Reported											
7	Therapy	HTA4_4	Yes	Chemotherapy	Not Reported	Unknown		Bone marrow		Initial Diagnosis											
8	Therapy	HTA4_5	Yes	Chemotherapy	Not Reported	Unknown		Bone marrow		Initial Diagnosis											
9	Therapy	HTA4_11	Yes	Chemotherapy	Not Reported	Unknown		Bone marrow		Initial Diagnosis											
10	Therapy	HTA4_8	Yes	Chemotherapy	Not Reported	Unknown		Bone marrow		Initial Diagnosis											
11	Therapy	HTA4_12	Yes	Chemotherapy	Not Reported	Unknown		Bone marrow		Initial Diagnosis											

Showing 1 to 10 of 30 entries

Validate Filled Metadata

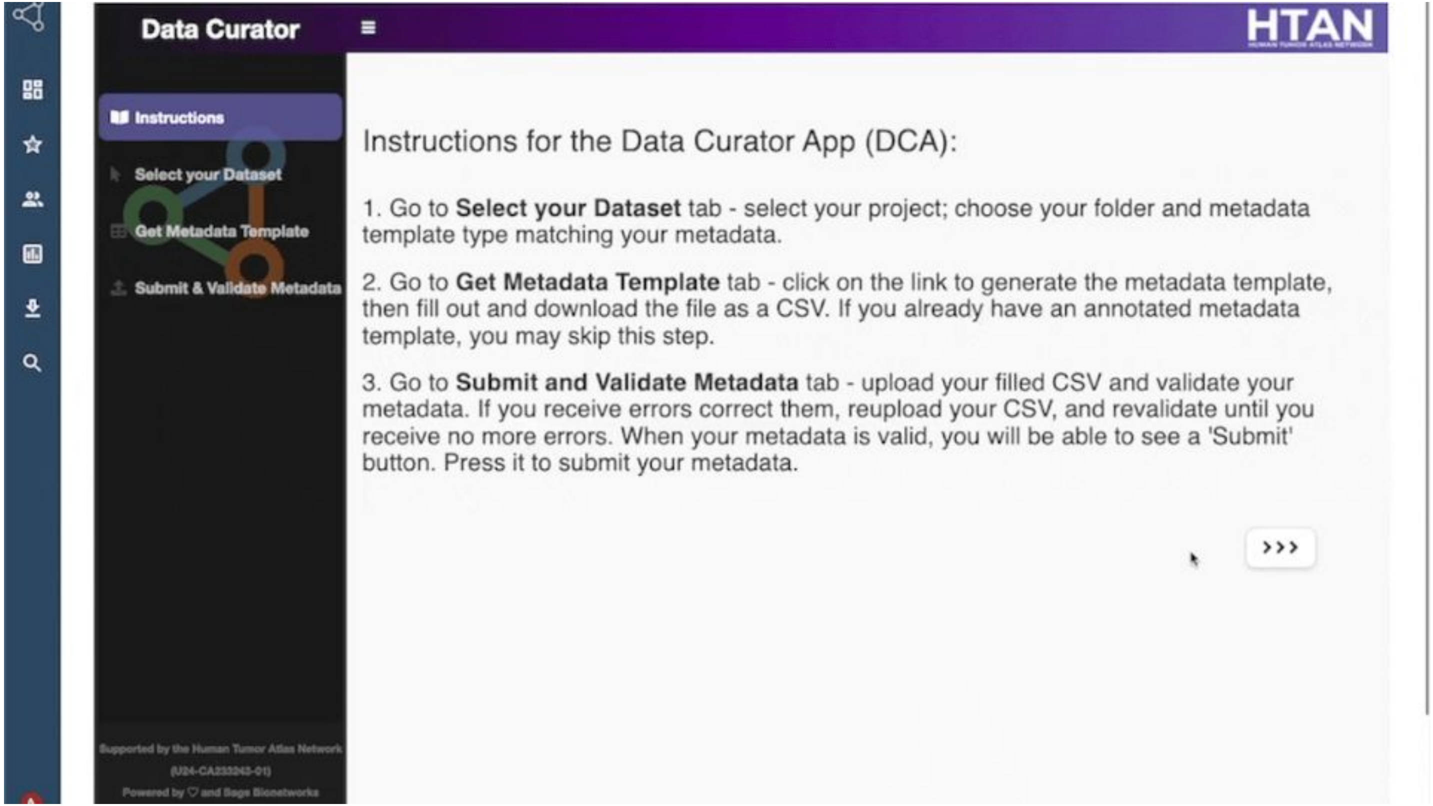
Your metadata is **invalid** !!!

The submitted metadata have 25 errors.

View all the error(s) highlighted in the preview table above

Row	Column	Value	Error
7, 8, 9, 10, 11, 12, 13, 14, 15 ... 31	Treatment Outcome	Unknown	'Unknown' is not one of ['Stable Disease', 'Very Good Partial Response', 'Not Reported', 'Complete Response', 'Progressive Disease', 'Treatment Stopped Due to Toxicity', 'No Measurable Disease', 'unknown', 'Persistent Disease', 'Partial Response', 'No Response', 'Mixed Response', 'Treatment Ongoing', '']

Metadata submission and validation

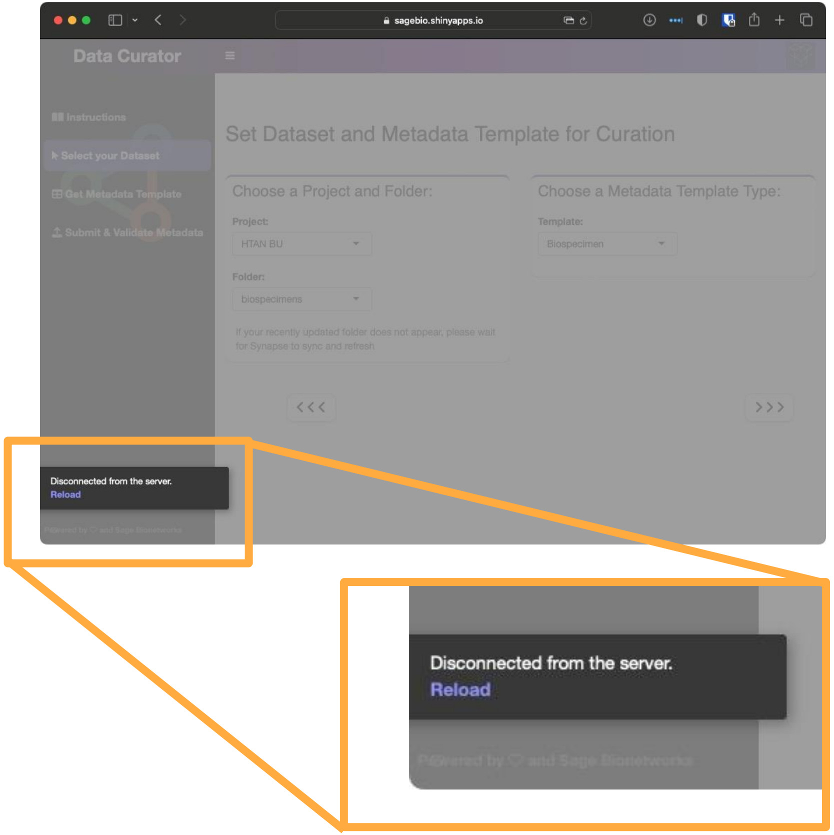


Instructions for the Data Curator App (DCA):

1. Go to **Select your Dataset** tab - select your project; choose your folder and metadata template type matching your metadata.
2. Go to **Get Metadata Template** tab - click on the link to generate the metadata template, then fill out and download the file as a CSV. If you already have an annotated metadata template, you may skip this step.
3. Go to **Submit and Validate Metadata** tab - upload your filled CSV and validate your metadata. If you receive errors correct them, reupload your CSV, and revalidate until you receive no more errors. When your metadata is valid, you will be able to see a 'Submit' button. Press it to submit your metadata.

HTAN DCA errors

- Disconnect error shows in the bottom left hand corner. (may be hidden behind transparency)
- Please open an HTAN service desk ticket (bit.ly/htan-help) and include the manifest that you were attempting to upload as an attachment.
- This helps us squash bugs and get your data submitted faster



Set Dataset and Metadata Template for Curation

Choose a Project and Folder:

Project: HTAN BU

Folder: biospecimens

If your recently updated folder does not appear, please wait for Synapse to sync and refresh

Choose a Metadata Template Type:

Template: Biospecimen

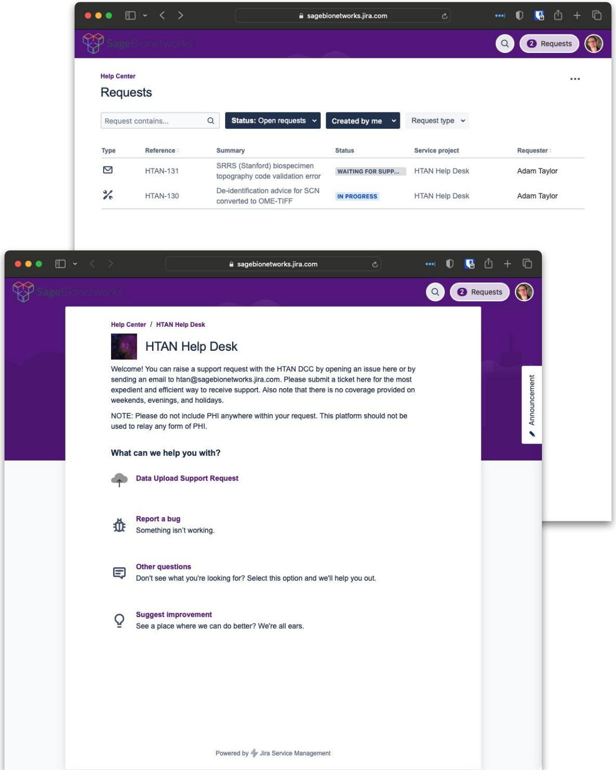
Disconnected from the server.
Reload

Useful DCC resources

Getting in touch with the DCC

(in rough order of preference)

- Create a HTAN DCC service desk ticket
    - bit.ly/htan-help or email **htan@sagebionetworks.jira.com**
- Message in the **#dcc-uploads-bu** channel on the HTAN Slack workspace



Requests

Type	Reference	Summary	Status	Service project	Requester
✉	HTAN-131	SRRS (Stanford) biospecimen topography code validation error	WAITING FOR SUPP...	HTAN Help Desk	Adam Taylor
✂	HTAN-130	De-identification advice for SCN converted to OME-TIFF	IN PROGRESS	HTAN Help Desk	Adam Taylor

HTAN Help Desk

Welcome! You can raise a support request with the HTAN DCC by opening an issue here or by sending an email to htan@sagebionetworks.jira.com. Please submit a ticket here for the most expedient and efficient way to receive support. Also note that there is no coverage provided on weekends, evenings, and holidays.

NOTE: Please do not include PHI anywhere within your request. This platform should not be used to relay any form of PHI.

What can we help you with?

Data Upload Support Request

Report a bug
Something isn't working.

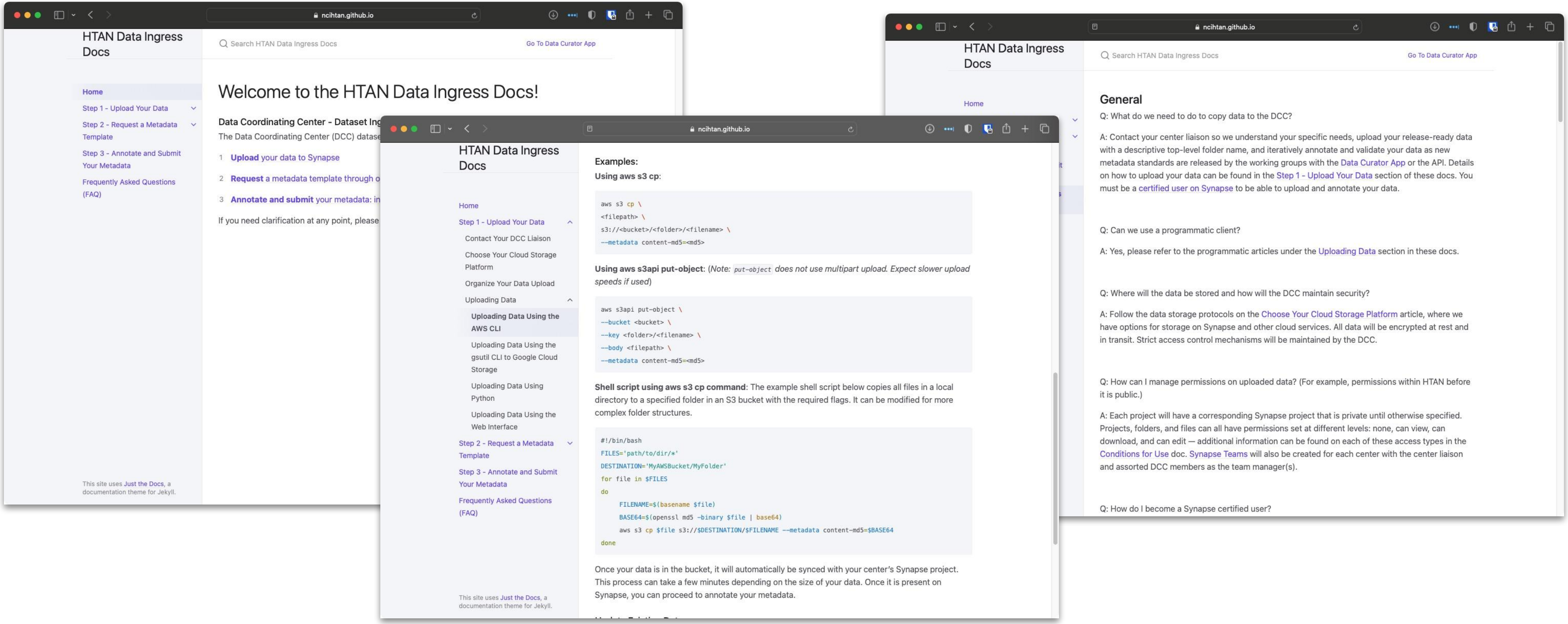
Other questions
Don't see what you're looking for? Select this option and we'll help you out.

Suggest improvement
See a place where we can do better? We're all ears.

Powered by Jira Service Management

HTAN Data Ingress Docs

ncihtan.github.io/HTAN-Data-Ingress-Docs/faq.html



Welcome to the HTAN Data Ingress Docs!

Data Coordinating Center - Dataset Ing

The Data Coordinating Center (DCC) datase

1. **Upload** your data to Synapse
2. **Request** a metadata template through o
3. **Annotate and submit** your metadata: in

If you need clarification at any point, please

Examples:

Using aws s3 cp:

```
aws s3 cp \
<filepath> \
s3://<bucket>/<folder>/<filename> \
--metadata content-md5=<md5>
```

Using aws s3api put-object: (*Note:* `put-object` *does not use multipart upload. Expect slower upload speeds if used*)

```
aws s3api put-object \
--bucket <bucket> \
--key <folder>/<filename> \
--body <filepath> \
--metadata content-md5=<md5>
```

Shell script using aws s3 cp command: The example shell script below copies all files in a local directory to a specified folder in an S3 bucket with the required flags. It can be modified for more complex folder structures.

```
#!/bin/bash
FILES='path/to/dir/*'
DESTINATION='MyAWSBucket/MyFolder'
for file in $FILES
do
    FILENAME=$(basename $file)
    BASE64=$(openssl md5 -binary $file | base64)
    aws s3 cp $file s3://$DESTINATION/$FILENAME --metadata content-md5=$BASE64
done
```

Once your data is in the bucket, it will automatically be synced with your center's Synapse project. This process can take a few minutes depending on the size of your data. Once it is present on Synapse, you can proceed to annotate your metadata.

General

Q: What do we need to do to copy data to the DCC?

A: Contact your center liaison so we understand your specific needs, upload your release-ready data with a descriptive top-level folder name, and iteratively annotate and validate your data as new metadata standards are released by the working groups with the Data Curator App or the API. Details on how to upload your data can be found in the Step 1 - Upload Your Data section of these docs. You must be a certified user on Synapse to be able to upload and annotate your data.

Q: Can we use a programmatic client?

A: Yes, please refer to the programmatic articles under the Uploading Data section in these docs.

Q: Where will the data be stored and how will the DCC maintain security?

A: Follow the data storage protocols on the Choose Your Cloud Storage Platform article, where we have options for storage on Synapse and other cloud services. All data will be encrypted at rest and in transit. Strict access control mechanisms will be maintained by the DCC.

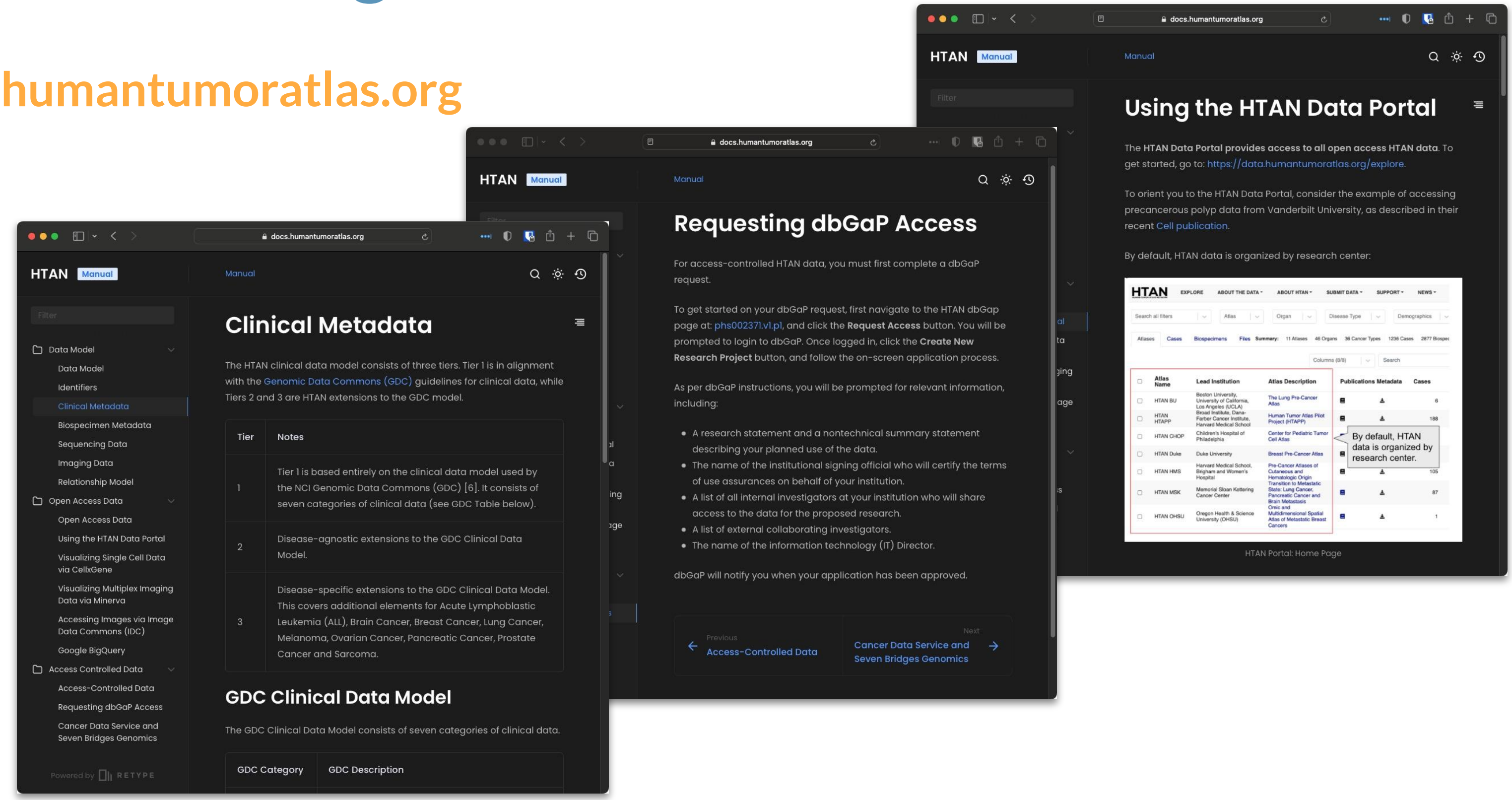
Q: How can I manage permissions on uploaded data? (For example, permissions within HTAN before it is public.)

A: Each project will have a corresponding Synapse project that is private until otherwise specified. Projects, folders, and files can all have permissions set at different levels: none, can view, can download, and can edit — additional information can be found on each of these access types in the Conditions for Use doc. Synapse Teams will also be created for each center with the center liaison and assorted DCC members as the team manager(s).

Q: How do I become a Synapse certified user?

HTAN Missing Manual

docs.humantumoratlas.org



Clinical Metadata

The HTAN clinical data model consists of three tiers. Tier 1 is in alignment with the Genomic Data Commons (GDC) guidelines for clinical data, while Tiers 2 and 3 are HTAN extensions to the GDC model.

Tier	Notes
1	Tier 1 is based entirely on the clinical data model used by the NCI Genomic Data Commons (GDC) [6]. It consists of seven categories of clinical data (see GDC Table below).
2	Disease-agnostic extensions to the GDC Clinical Data Model.
3	Disease-specific extensions to the GDC Clinical Data Model. This covers additional elements for Acute Lymphoblastic Leukemia (ALL), Brain Cancer, Breast Cancer, Lung Cancer, Melanoma, Ovarian Cancer, Pancreatic Cancer, Prostate Cancer and Sarcoma.

GDC Clinical Data Model

The GDC Clinical Data Model consists of seven categories of clinical data.

GDC Category	GDC Description

Requesting dbGaP Access

For access-controlled HTAN data, you must first complete a dbGaP request.

To get started on your dbGaP request, first navigate to the HTAN dbGap page at: phs002371.v1.p1, and click the **Request Access** button. You will be prompted to login to dbGaP. Once logged in, click the **Create New Research Project** button, and follow the on-screen application process.

As per dbGaP instructions, you will be prompted for relevant information, including:

- A research statement and a nontechnical summary statement describing your planned use of the data.
- The name of the institutional signing official who will certify the terms of use assurances on behalf of your institution.
- A list of all internal investigators at your institution who will share access to the data for the proposed research.
- A list of external collaborating investigators.
- The name of the information technology (IT) Director.

dbGaP will notify you when your application has been approved.

Using the HTAN Data Portal

The **HTAN Data Portal provides access to all open access HTAN data**. To get started, go to: https://data.humantumoratlas.org/explore.

To orient you to the HTAN Data Portal, consider the example of accessing precancerous polyp data from Vanderbilt University, as described in their recent Cell publication.

By default, HTAN data is organized by research center:

☐	Atlas Name	Lead Institution	Atlas Description	Publications	Metadata	Cases
☐	HTAN BU	Boston University, University of California, Los Angeles (UCLA)	The Lung Pre-Cancer Atlas	■	⤓	6
☐	HTAN HTAPP	Broad Institute, Dana-Farber Cancer Institute, Harvard Medical School	Human Tumor Atlas Pilot Project (HTAPP)	■	⤓	188
☐	HTAN CHOP	Children's Hospital of Philadelphia	Center for Pediatric Tumor Cell Atlas			
☐	HTAN Duke	Duke University	Breast Pre-Cancer Atlas	■		
☐	HTAN HMS	Harvard Medical School, Brigham and Women's Hospital	Pre-Cancer Atlases of Cutaneous and Hematologic Origin	■	⤓	105
☐	HTAN MSK	Memorial Sloan Kettering Cancer Center	Transition to Metastatic State: Lung Cancer, Pancreatic Cancer and Brain Metastasis	■	⤓	87
☐	HTAN OHSU	Oregon Health & Science University (OHSU)	Omic and Multidimensional Spatial Atlas of Metastatic Breast Cancers	■	⤓	1

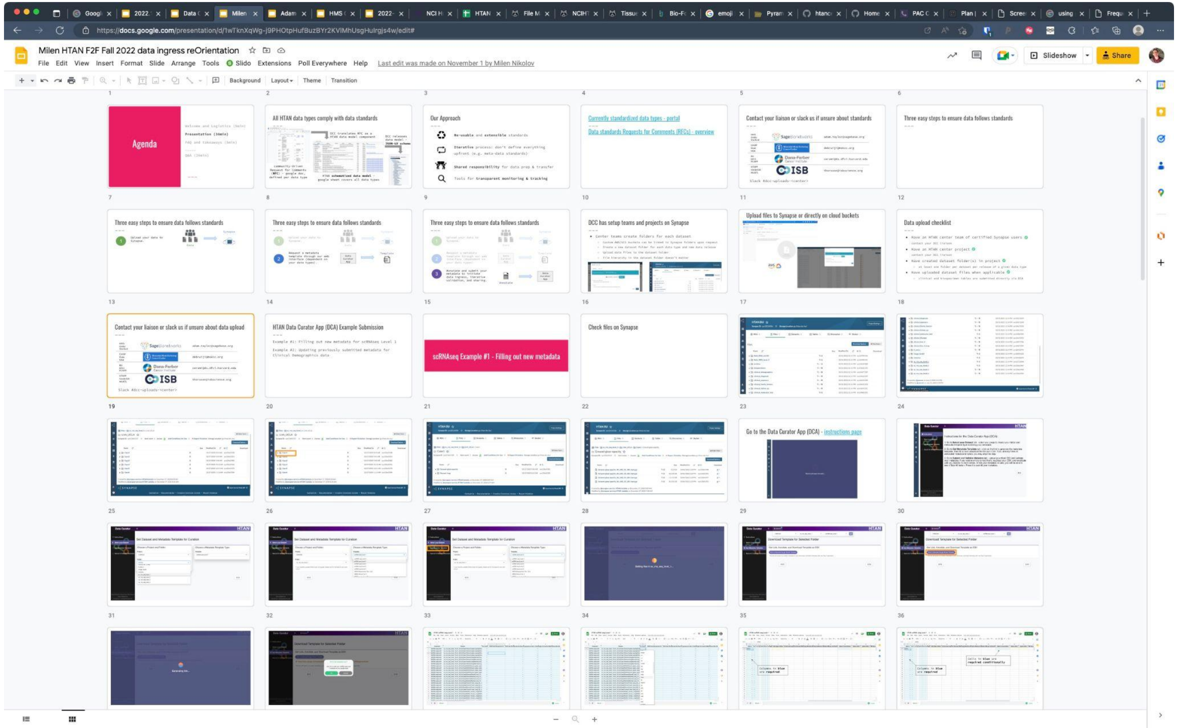
By default, HTAN data is organized by research center.

HTAN Portal: Home Page

DCC re-orientation at Fall F2F

Slides on Synapse:

syn45352565



https://docs.google.com/presentation/d/1wTknXqWg-j9PHOtpHufBuzBYr2KVIMhUsgHulrgjs4w/edit#

Milen HTAN F2F Fall 2022 data ingress reOrientation

Last edit was made on November 1 by Milen Nikolov

1 Agenda

2 All HTAN data types comply with data standards

3 Our Approach

- Re-usable and extensible standards
- Iterative process: don't define everything upfront (e.g. meta-data standards)
- Shared responsibility for data prep & transfer
- Tools for transparent monitoring & tracking

4 Currently standardized data types - portal

Data standards Requests for Comments (RFCs) - overview

5 Contact your liaison or slack us if unsure about standards

6 Three easy steps to ensure data follows standards

7 Three easy steps to ensure data follows standards

8 Three easy steps to ensure data follows standards

9 Three easy steps to ensure data follows standards

10 DCC has setup teams and projects on Synapse

11 Upload files to Synapse or directly on cloud buckets

12 Data upload checklist

13 Contact your liaison or slack us if unsure about data upload

14 HTAN Data Curator App (DCA) Example Submission

15 scRNASeq Example #1 - Filling out new metadata

16 Check files on Synapse

29 Go to the Data Curator App (DCA) - instructions page

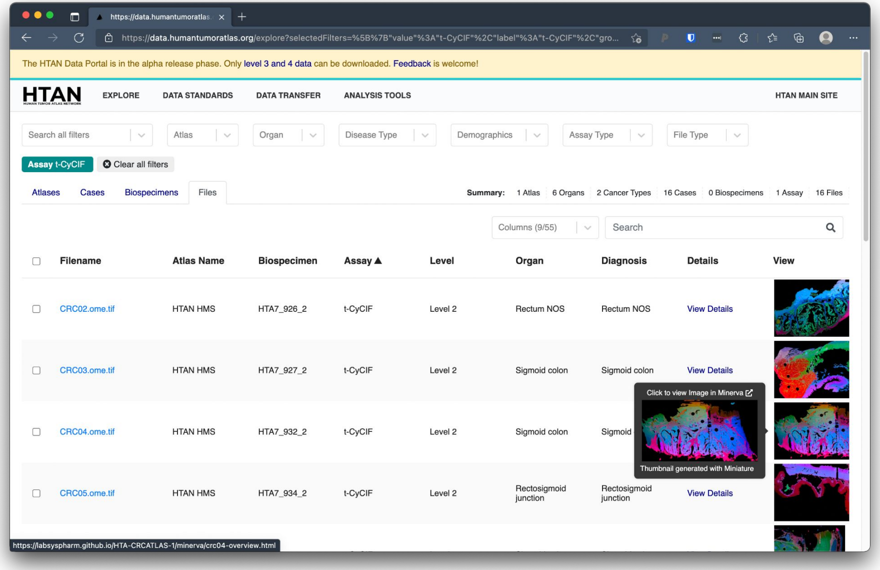
HTAN Data Portal and data destinations

Ino de Bruijn, MSK

HTAN Portal

data.humantumoratlas.org

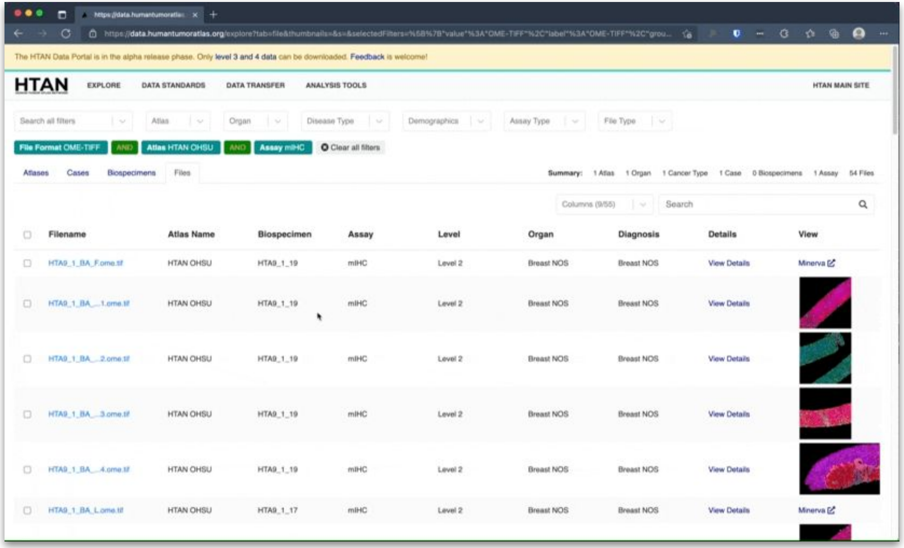
Metadata filtering, thumbnails and linked image visualisation



The HTAN Data Portal is in the alpha release phase. Only level 3 and 4 data can be downloaded. Feedback is welcome!

Summary: 1 Atlas 6 Organs 2 Cancer Types 16 Cases 0 Biospecimens 1 Assay 16 Files

☐	Filename	Atlas Name	Biospecimen	Assay ▲	Level	Organ	Diagnosis	Details	View
☐	CRC02.ome.tif	HTAN HMS	HTA7_926_2	t-CyCIF	Level 2	Rectum NOS	Rectum NOS	View Details	
☐	CRC03.ome.tif	HTAN HMS	HTA7_927_2	t-CyCIF	Level 2	Sigmoid colon	Sigmoid colon	View Details	
☐	CRC04.ome.tif	HTAN HMS	HTA7_932_2	t-CyCIF	Level 2	Sigmoid colon	Sigmoid		
☐	CRC05.ome.tif	HTAN HMS	HTA7_934_2	t-CyCIF	Level 2	Rectosigmoid junction	Rectosigmoid junction	View Details	



The HTAN Data Portal is in the alpha release phase. Only level 3 and 4 data can be downloaded. Feedback is welcome!

Summary: 1 Atlas 1 Organ 1 Cancer Type 1 Case 0 Biospecimens 1 Assay 54 Files

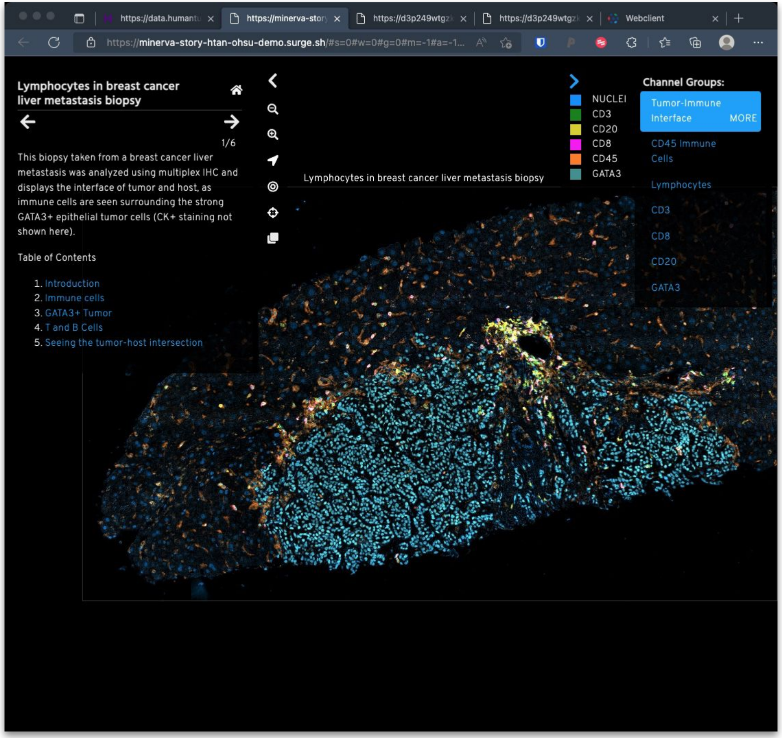
☐	Filename	Atlas Name	Biospecimen	Assay	Level	Organ	Diagnosis	Details	View
☐	HTA9_1_BA_F.ome.tif	HTAN OHSU	HTA9_1_19	mIHC	Level 2	Breast NOS	Breast NOS	View Details	Minerva
☐	HTA9_1_BA_...1.ome.tif	HTAN OHSU	HTA9_1_19	mIHC	Level 2	Breast NOS	Breast NOS	View Details	
☐	HTA9_1_BA_...2.ome.tif	HTAN OHSU	HTA9_1_19	mIHC	Level 2	Breast NOS	Breast NOS	View Details	
☐	HTA9_1_BA_...3.ome.tif	HTAN OHSU	HTA9_1_19	mIHC	Level 2	Breast NOS	Breast NOS	View Details	
☐	HTA9_1_BA_...4.ome.tif	HTAN OHSU	HTA9_1_19	mIHC	Level 2	Breast NOS	Breast NOS	View Details	
☐	HTA9_1_BA_L.ome.tif	HTAN OHSU	HTA9_1_17	mIHC	Level 2	Breast NOS	Breast NOS	View Details	Minerva

Jeremy Muhlich, HMS

Minerva stories

Pre-rendered channel overlay groups with waypoints and annotation sharing. Wide potential for overlays and supplementary graphs etc

Manually curated



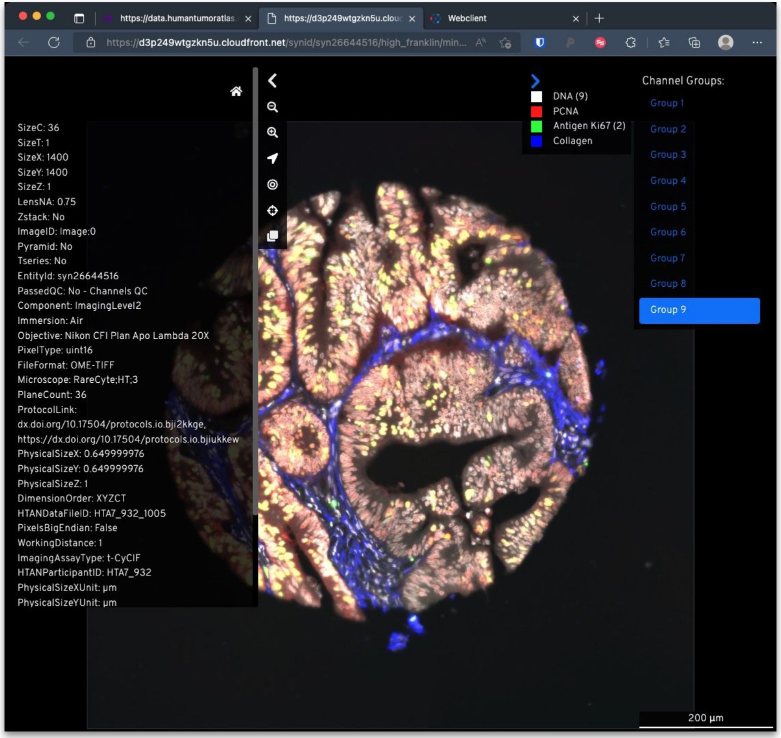
Lymphocytes in breast cancer liver metastasis biopsy

This biopsy taken from a breast cancer liver metastasis was analyzed using multiplex IHC and displays the interface of tumor and host, as immune cells are seen surrounding the strong GATA3+ epithelial tumor cells (CK+ staining not shown here).

Table of Contents

1. Introduction
2. Immune cells
3. GATA3+ Tumor
4. T and B Cells
5. Seeing the tumor-host intersection

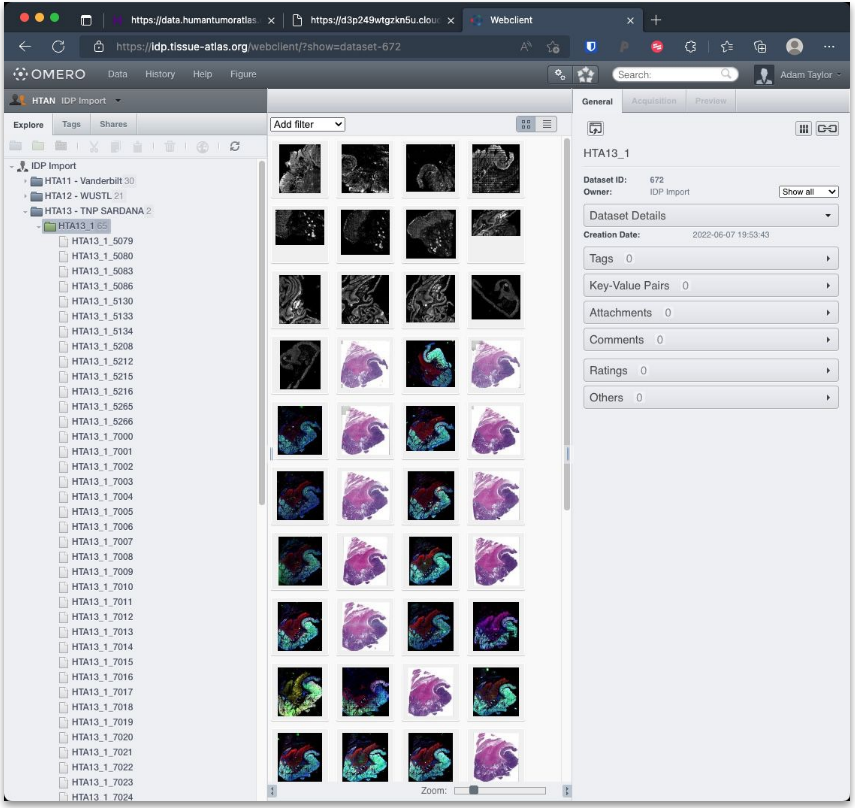
Automatic default (Auto-Minerva)



SizeC: 36
SizeT: 1
SizeX: 1400
SizeY: 1400
SizeZ: 1
LensNA: 0.75
Zstack: No
ImageID: Image:0
Pyramid: No
Tseries: No
EntityId: syn26644516
PassedQC: No - Channels QC
Component: ImagingLevel2
Immersion: Air
Objective: Nikon CFI Plan Apo Lambda 20X
PixelType: uint16
FileFormat: OME-TIFF
Microscope: RareCyte;HT;3
PlaneCount: 36
ProtocolLink:
dx.doi.org/10.17504/protocols.io.bji2kkge,
https://dx.doi.org/10.17504/protocols.io.bjiukkew
PhysicalSizeX: 0.649999976
PhysicalSizeY: 0.649999976
PhysicalSizeZ: 1
DimensionOrder: XYZCT
HTANDataFileID: HTA7_932_1005
PixelsBigEndian: False
WorkingDistance: 1
ImagingAssayType: t-CyCIF
HTANParticipantID: HTA7_932
PhysicalSizeXUnit: µm
PhysicalSizeYUnit: µm

HTAN IDP OMERO instance

- Synapse authentication
- HTAN Release 1 and 2 images
- Also used by TNP TMA and TNP SARDANA for initial staging
- Auto-minerva preset rendering



https://idp.tissue-atlas.org/webclient/?show=dataset-672

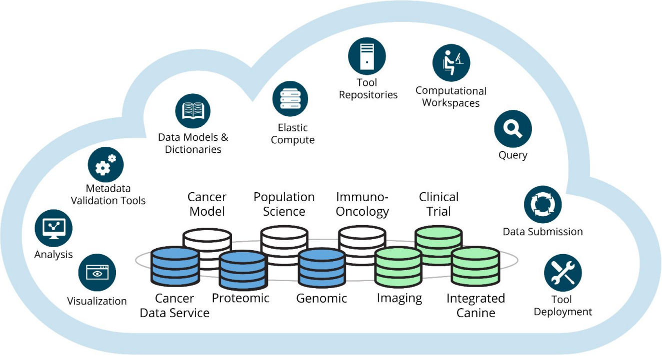
HTA13_1

Dataset ID: 672
Owner: IDP Import

Creation Date: 2022-06-07 19:53:43

HTAN Data Destinations

NATIONAL CANCER INSTITUTE
Cancer Research Data Commons



SYNAPSE

CANCER GENOMICS CLOUD
SEVEN BRIDGES

A RESOURCE OF THE NCI CANCER RESEARCH DATA COMMONS

ISB-CGC

Cancer Gateway in the Cloud

Access, Explore and Analyze Large-Scale Cancer Data Through the Google Cloud

Google Big Query

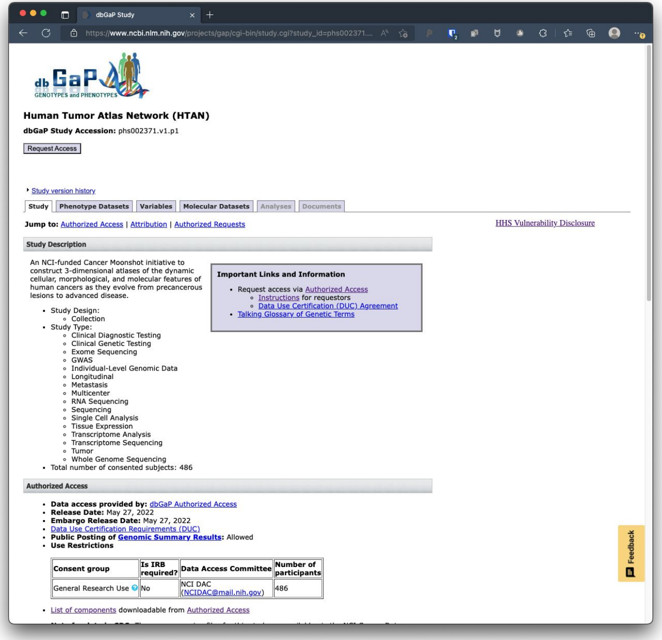
Level 1 and 2 Sequencing ➡
NIH CRDC CDS
authenticated through dbGaP
available through Seven Bridges Genomics Cloud

Level 3 and 4 Sequencing ➡
Available through Synapse, and through Google BigQuery at ISB-CGC (June 2022)

Imaging ➡
CDS and Imaging Data Commons

Seven Bridges Cancer Genomics Cloud

dbGaP Study Accession: phs002371.v1.p1



Human Tumor Atlas Network (HTAN)

dbGaP Study Accession: phs002371.v1.p1

Jump to: Authorized Access | Attribution | Authorized Requests

Study Description

An NCI-funded Cancer Moonshot initiative to construct 3-dimensional atlases of the dynamic cellular, morphological, and molecular features of human cancers as they evolve from precancerous lesions to advanced disease.

- Study Design:
    - Collection
- Study Type:
    - Clinical Diagnostic Testing
    - Clinical Genetic Testing
    - Exome Sequencing
    - GWAS
    - Individual-Level Genomic Data
    - Longitudinal
    - Metastasis
    - Multicenter
    - RNA Sequencing
    - Sequencing
    - Single Cell Analysis
    - Tissue Expression
    - Transcriptome Analysis
    - Transcriptome Sequencing
    - Tumor
    - Whole Genome Sequencing
- Total number of consented subjects: 486

Important Links and Information

- Request access via Authorized Access
    - Instructions for requestors
    - Data Use Certification (DUC) Agreement
- Talking Glossary of Genetic Terms

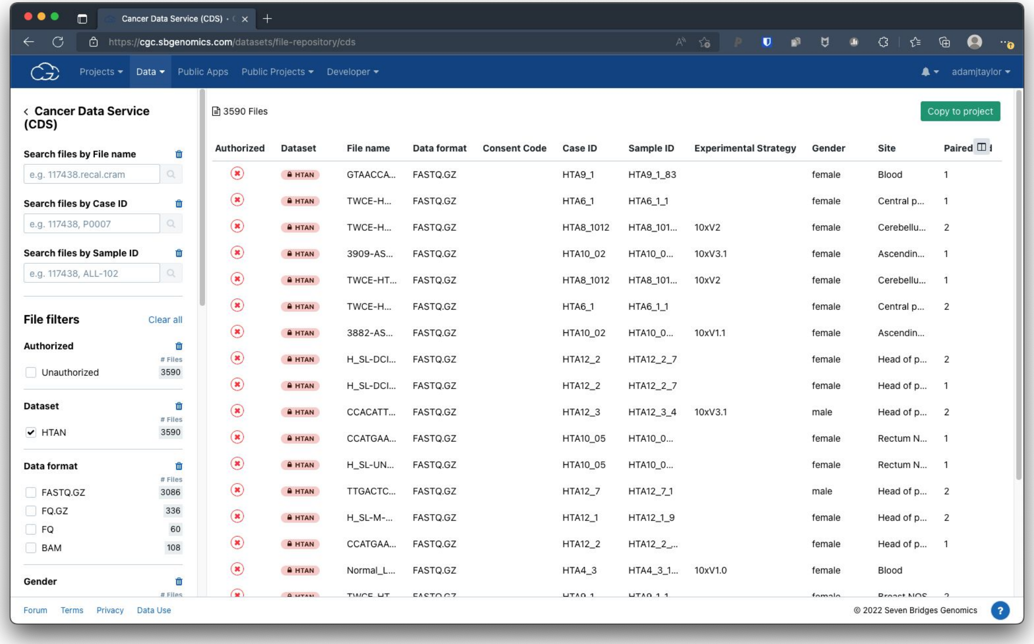
Authorized Access

- **Data access provided by:** dbGaP Authorized Access
- **Release Date:** May 27, 2022
- **Embargo Release Date:** May 27, 2022
- Data Use Certification Requirements (DUC)
- **Public Posting of Genomic Summary Results:** Allowed
- **Use Restrictions**

Consent group	Is IRB required?	Data Access Committee	Number of participants
General Research Use	No	NCI DAC (NCIDAC@mail.nih.gov)	486

- List of components downloadable from Authorized Access

Visible in SB CGC though CDS Explorer page



Cancer Data Service (CDS)

Search files by File name
e.g. 117438.recal.cram

Search files by Case ID
e.g. 117438, P0007

Search files by Sample ID
e.g. 117438, ALL-102

File filters Clear all

Authorized — # Files
☐ Unauthorized 3590

Dataset — # Files
☑ HTAN 3590

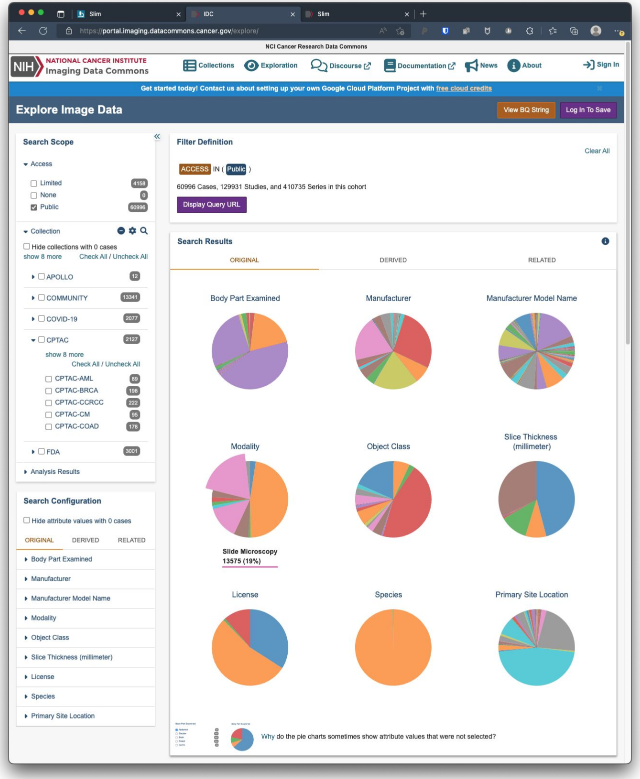
Data format — # Files
☐ FASTQ.GZ 3086
☐ FQ.GZ 336
☐ FQ 60
☐ BAM 108

Gender

3590 Files

Authorized	Dataset	File name	Data format	Consent Code	Case ID	Sample ID	Experimental Strategy	Gender	Site	Paired
☒	HTAN	GTAACCA...	FASTQ.GZ		HTA9_1	HTA9_1_83		female	Blood	1
☒	HTAN	TWCE-H...	FASTQ.GZ		HTA6_1	HTA6_1_1		female	Central p...	1
☒	HTAN	TWCE-H...	FASTQ.GZ		HTA8_1012	HTA8_101...	10xV2	female	Cerebellu...	2
☒	HTAN	3909-AS...	FASTQ.GZ		HTA10_02	HTA10_0...	10xV3.1	female	Ascendin...	1
☒	HTAN	TWCE-HT...	FASTQ.GZ		HTA8_1012	HTA8_101...	10xV2	female	Cerebellu...	1
☒	HTAN	TWCE-H...	FASTQ.GZ		HTA6_1	HTA6_1_1		female	Central p...	2
☒	HTAN	3882-AS...	FASTQ.GZ		HTA10_02	HTA10_0...	10xV1.1	female	Ascendin...	
☒	HTAN	H_SL-DCI...	FASTQ.GZ		HTA12_2	HTA12_2_7		female	Head of p...	2
☒	HTAN	H_SL-DCI...	FASTQ.GZ		HTA12_2	HTA12_2_7		female	Head of p...	1
☒	HTAN	CCACATT...	FASTQ.GZ		HTA12_3	HTA12_3_4	10xV3.1	male	Head of p...	2
☒	HTAN	CCATGAA...	FASTQ.GZ		HTA10_05	HTA10_0...		female	Rectum N...	1
☒	HTAN	H_SL-UN...	FASTQ.GZ		HTA10_05	HTA10_0...		female	Rectum N...	1
☒	HTAN	TTGACTC...	FASTQ.GZ		HTA12_7	HTA12_7_1		male	Head of p...	2
☒	HTAN	H_SL-M-...	FASTQ.GZ		HTA12_1	HTA12_1_9		female	Head of p...	2
☒	HTAN	CCATGAA...	FASTQ.GZ		HTA12_2	HTA12_2_...		female	Head of p...	1
☒	HTAN	Normal_L...	FASTQ.GZ		HTA4_3	HTA4_3_1...	10xV1.0	female	Blood	

Imaging data also going to Imaging Data Commons



Get started today! Contact us about setting up your own Google Cloud Platform Project with free cloud credits

☐ Limited 4158
☐ None 0
☑ Public 60996

☐ Hide collections with 0 cases

☐ APOLLO 12
☐ COMMUNITY 13341
☐ COVID-19 2077
☐ CPTAC 2127
☐ CPTAC-AML 89
☐ CPTAC-BRCA 198
☐ CPTAC-CCRCC 222
☐ CPTAC-CM 95
☐ CPTAC-COAD 178
☐ FDA 3001

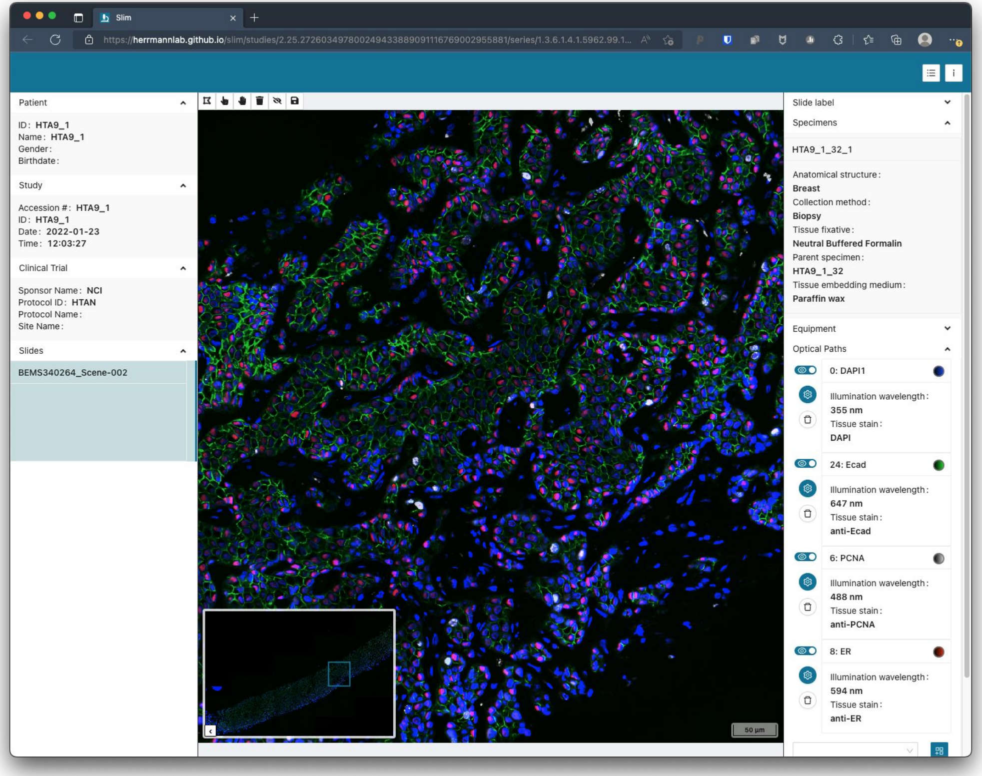
☐ Hide attribute values with 0 cases

- Body Part Examined
- Manufacturer
- Manufacturer Model Name
- Modality
- Object Class
- Slice Thickness (millimeter)
- License
- Species
- Primary Site Location

ACCESS IN (Public)

60996 Cases, 129931 Studies, and 410735 Series in this cohort

Why do the pie charts sometimes show attribute values that were not selected?



ID: HTA9_1
Name: HTA9_1
Gender:
Birthdate:

Accession #: HTA9_1
ID: HTA9_1
Date: 2022-01-23
Time: 12:03:27

Sponsor Name: NCI
Protocol ID: HTAN
Protocol Name:
Site Name:

BEMS340264_Scene-002

HTA9_1_32_1

Anatomical structure: Breast
Collection method: Biopsy
Tissue fixative: Neutral Buffered Formalin
Parent specimen: HTA9_1_32
Tissue embedding medium: Paraffin wax

0: DAPI1
Illumination wavelength: 355 nm
Tissue stain: DAPI

24: Ecad
Illumination wavelength: 647 nm
Tissue stain: anti-Ecad

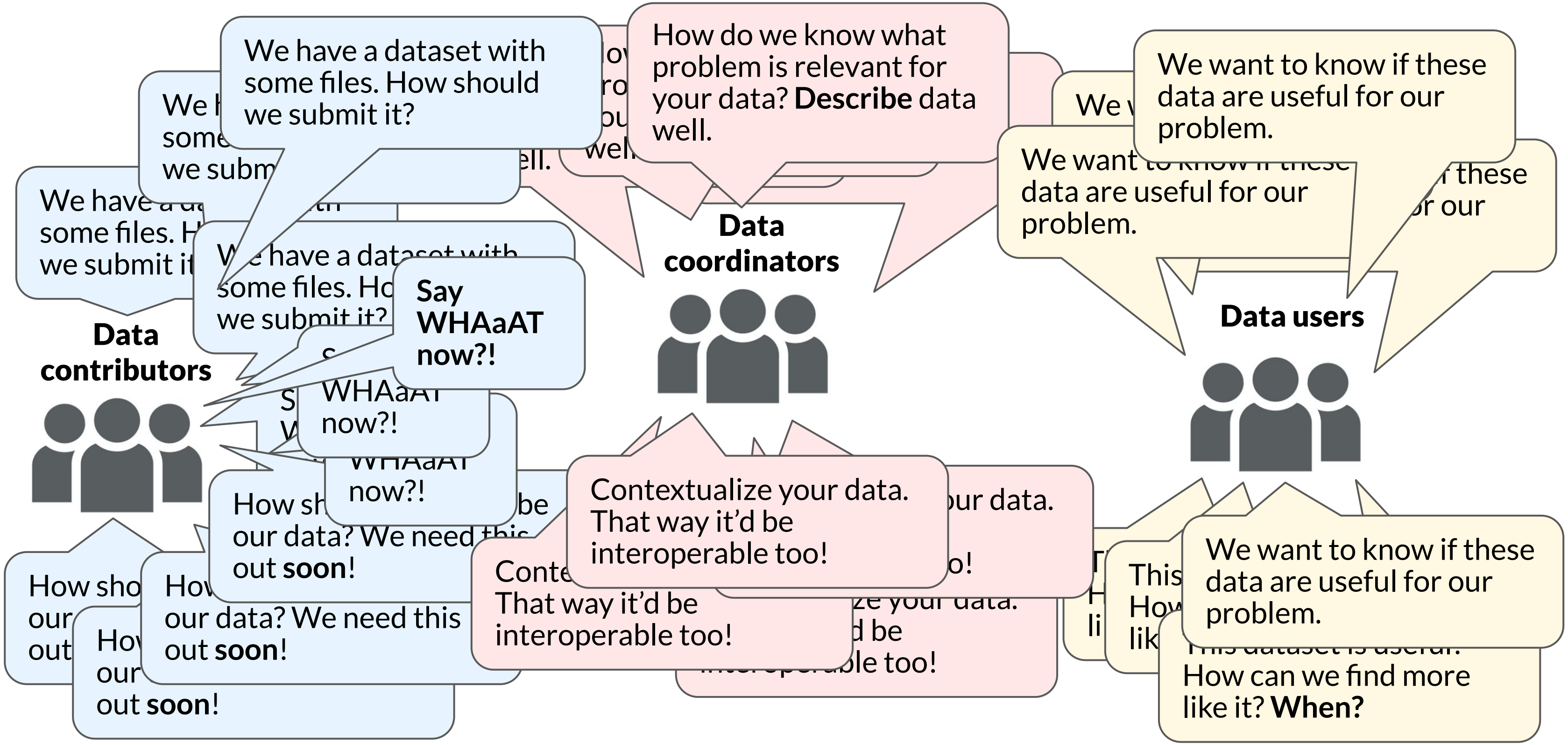
6: PCNA
Illumination wavelength: 488 nm
Tissue stain: anti-PCNA

8: ER
Illumination wavelength: 594 nm
Tissue stain: anti-ER

Final thoughts

Enabling FAIR data requires COMMUNICATION

at SCALE…



Data contributors

We have a dataset with some files. How should we submit it?

Say WHAaAT now?!

How should we … our data? We need this out **soon**!

Data coordinators

How do we know what problem is relevant for your data? **Describe** data well.

Contextualize your data. That way it'd be interoperable too!

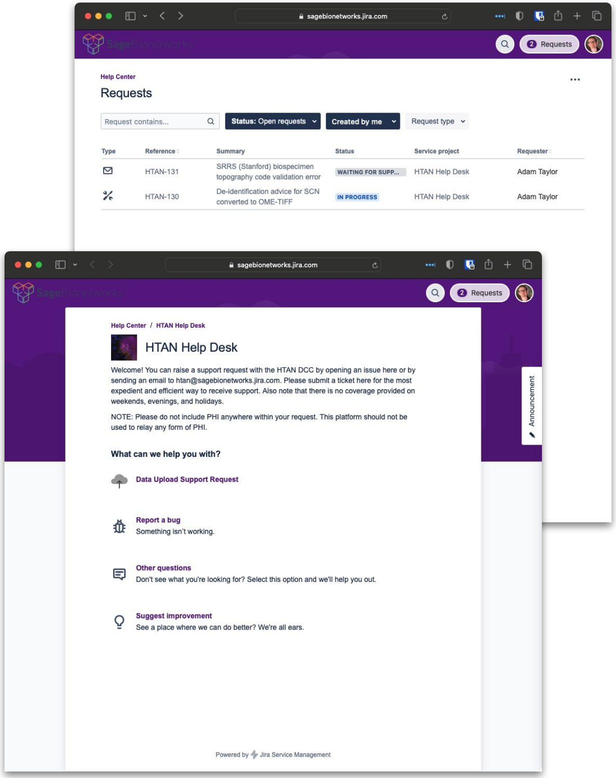
Data users

We want to know if these data are useful for our problem.

This dataset is useful. How can we find more like it? **When?**

Let's keep talking

- Check that we have templates for your data types - if not let us know.
- Let us know your publication plans and timelines
- Share data and metadata early
- Create a HTAN DCC service desk ticket so we can track issues bit.ly/htan-help or email **htan@sagebionetworks.jira.com**



Requests

Type	Reference	Summary	Status	Service project	Requester
✉	HTAN-131	SRRS (Stanford) biospecimen topography code validation error	WAITING FOR SUPP...	HTAN Help Desk	Adam Taylor
✂	HTAN-130	De-identification advice for SCN converted to OME-TIFF	IN PROGRESS	HTAN Help Desk	Adam Taylor

HTAN Help Desk

Welcome! You can raise a support request with the HTAN DCC by opening an issue here or by sending an email to htan@sagebionetworks.jira.com. Please submit a ticket here for the most expedient and efficient way to receive support. Also note that there is no coverage provided on weekends, evenings, and holidays.

NOTE: Please do not include PHI anywhere within your request. This platform should not be used to relay any form of PHI.

What can we help you with?

Data Upload Support Request

Report a bug
Something isn't working.

Other questions
Don't see what you're looking for? Select this option and we'll help you out.

Suggest improvement
See a place where we can do better? We're all ears.

Powered by Jira Service Management


James


Milen


Adam


Ashley


Elivra


Vésteinn


Ethan


Rochelle


Niki


Ino


Bill


Clarisse


Dave


Dar’ya

Aaron Lisman

Bruno Grande

Mialy DeFelice

Gianna Jordan

Onur Sumer

Xiang Li

Pierrette Lo

Tiarah Jeudi

Thank you!

adam.taylor@sagebase.org | bit.ly/htan-help