

Standardizing Zero-Tolerance De-identification



Michael W. Rutherford, M.S.

Instructor / PhD Candidate

Department of Biomedical Informatics

University of Arkansas for Medical Sciences









TCIA Collections



Data Collection

Specialized de-identification processes and tools support data collection, curation, and deidentification

Specialized Teams

Data Hosting and Query Research-focused website GUI and programmatic interfaces

Radiology

- 70,420 subjects
- 169,567 studies
- 79,682,647 images
- 31.4 TB of data
- Modalities
 - (CT / MR / PET / MG / US / SR / RT)
- File types:
 - (DICOM / NIFTI / NRRD)
- 250 TB downloads / month

Pathology

- 7,147 subjects
- 219,526 images
- 12.38 TB of data
- Modalities:
 - (WSI / microarray / IFC photomicrograph / raster)
- File types:
 - (SVS / TIFF / NDPI / MRXS / JPG, HDF5, BMP)
- 25 TB download / month

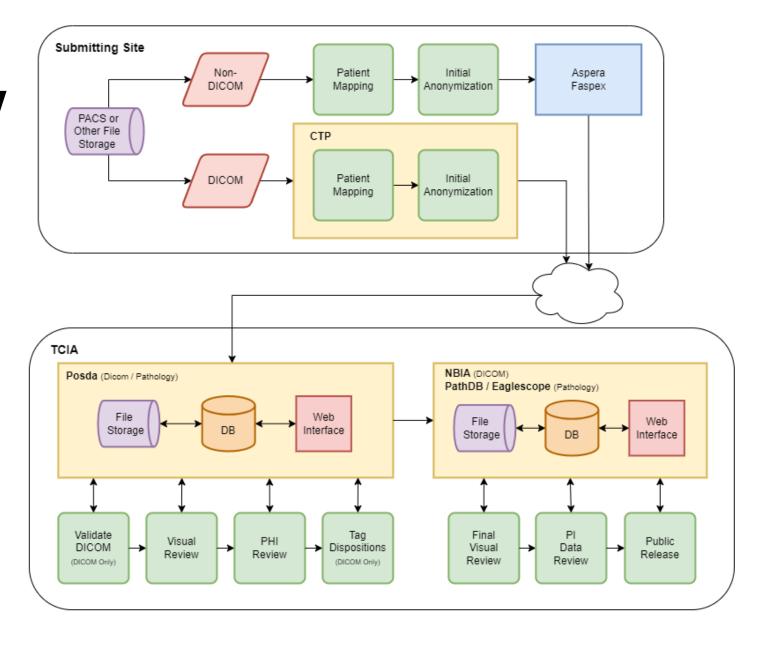






TCIA Workflow

- Submission
 - CTP
 - Aspera Faspex
- Curation
 - Posda
- Publication
 - NBIA
 - PathDB / Eaglescope











Submission

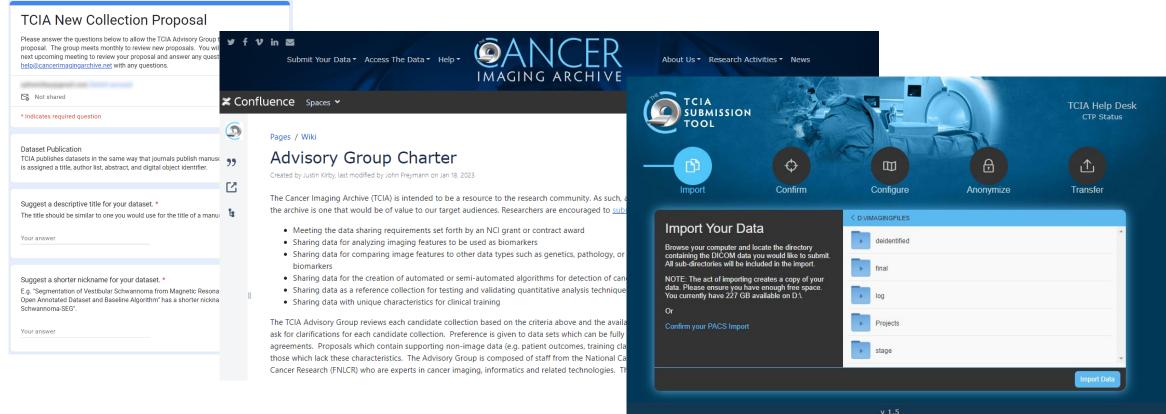
Submit **Proposal**



Monthly review by TCIA **Advisory Group**



TCIA **Submission** and **Curation** Support



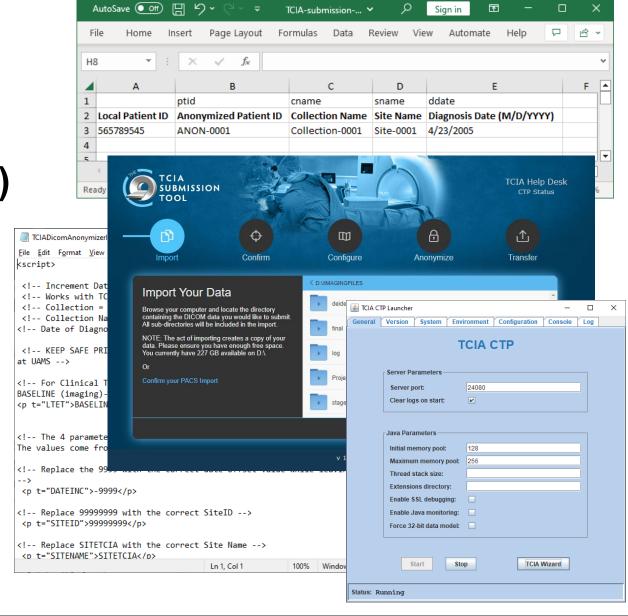






Submission Tools

- Clinical Trials Processor (CTP)
 - TCIA Submission Wizard
 - Import Files
 - Map Patients
 - Anonymize Files
 - Transfer Files
- Aspera Faspex
- Box









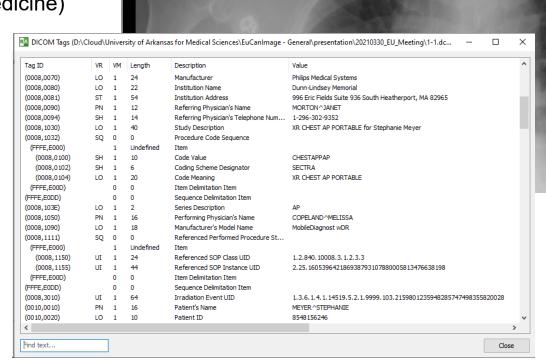


Medical Imaging

DICOM

(Digital Imaging and Communications in Medicine)

- Metadata (Header)
- Pixel Data
 - Content
 - Implants
 - Recognizable Features
 - Faces
 - Embedded
 - Burned-in Text
 - Watermark



MEYER STEPHANIE [F] 02.25.2012 DOB: 07.16.1953

Rutherford M, Mun SK, Levine B, Bennett W, Smith K, Farmer P, Jarosz Q, Wagner U, Freyman J, Blake G, Tarbox L, Farahani K, Prior F. *A DICOM dataset for evaluation of medical image de-identification*. Sci Data. 2021 Jul 16;8(1):183.

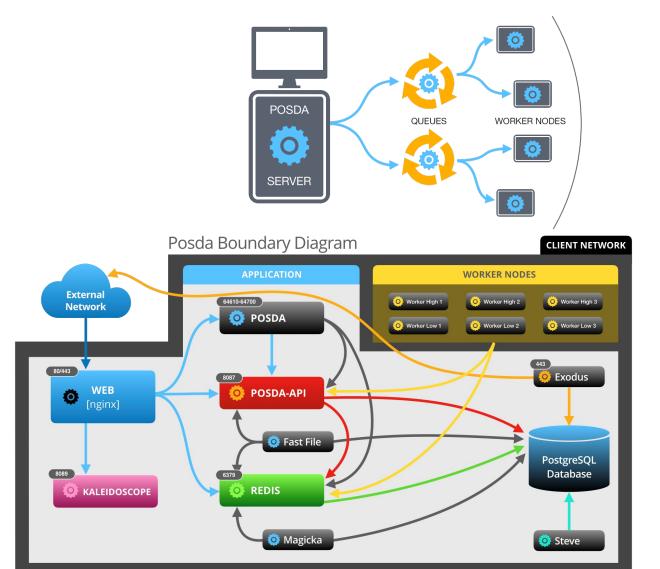






Posda

- Timepoint Architecture
- Worker Nodes
- Activity Workflows
- Visualization Tools
- Multi-format
- Multi-language









DICOM Conformance

- DICOM Validator (DCIODVFY)
- DICOM Std. PS3.15 E.1
- Basic Confidentiality Profile
 - Retain Safe Private
 - Retain UIDs
 - Retain Device Identity
 - Retain Institution Identity
 - Retain Patient Characteristics
 - Retain Longitudinal With Full Dates
 - Retain Longitudinal With Modified Dates
 - Clean Descriptors
 - Clean Structured Content.
 - Clean **Graphics**
 - Clean Pixel Data
 - Clean Recognizable Visual Features

replace with a non-zero length value that may be a dummy value and consistent with the VR
replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
remove
keep (unchanged for non-sequence attributes, cleaned for sequences)
clean, that is replace with values of similar meaning known not to contain identifying information and consistent with the VR
replace with a non-zero length UID that is internally consistent within a set of Instances
Z unless D is required to maintain IOD conformance (Type 2 versus Type 1)
X unless Z is required to maintain IOD conformance (Type 3 versus Type 2)
X unless D is required to maintain IOD conformance (Type 3 versus Type 1)
X unless Z or D is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1)
X unless Z or replacement of contained instance UIDs (U) is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1 sequences containing UID references)

attribute	tag	basic	priv	uid	dev	inst	pat	full	dates	desc	str	graph
Accession Number	<(0008,0050)>	Z										
Admitting Diagnoses Description	<(0008,1080)>	X								С		
Allergies	<(0010,2110)>	X					С			С		1
Curve Data	<(50xx,xxxx)>	X										С
Institution Address	<(0008,0081)>	X				K						
Institution Name	<(0008,0080)>	X/Z/D				K						
Overlay Data	<(60xx,3000)>	X										С
Patient's Age	<(0010,1010)>	X					K					
Patient's Birth Date	<(0010,0030)>	Z										
Patient's Name	<(0010,0010)>	Z										
Patient's Sex	<(0010,0040)>	Z					K					
Patient's Size	<(0010,1020)>	X					K					
Patient's Weight	<(0010,1030)>	X					K					
Private attributes	<(gggg,eeee)>	X	С									
	where gggg is odd											
Radiopharmaceutical Start DateTime	<(0018,1078)>	X						K	С			
Radiopharmaceutical Start Time	<(0018,1072)>	X						K	С			
Series Date	<(0008,0021)>	X/D						K	С			
Series Description	<(0008,103E)>	X								С		
SOP Instance UID	<(0008,0018)>	U		K								
Study Date	<(0008,0020)>	Z						K	С			
Study Description	<(0008,1030)>	X								С		







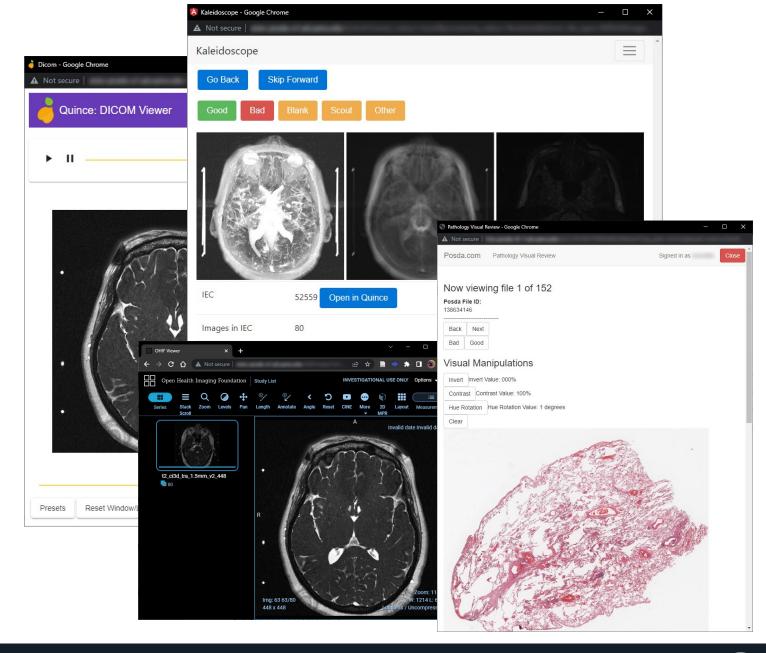
Visual Review

Kaleidoscope

- Mark viability (good/bad).
- Review series holistically (all images at same time)
- Max, Avg, Min projections

Pathology Visual Reviewer

- Invert, Contrast, Hue Rotation
- Other viewers (for deeper review and cleaning)
 - Quince, OHIF (ohif.org)
 - We use various external viewers as needed for testing purposes.

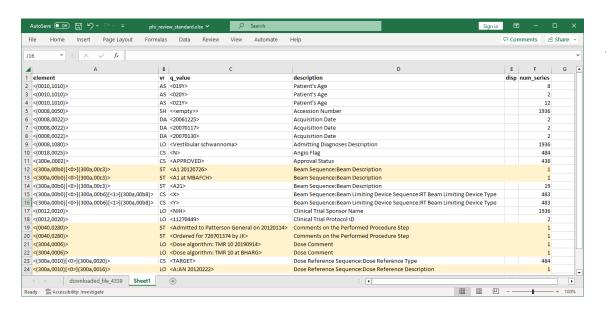








PHI Review



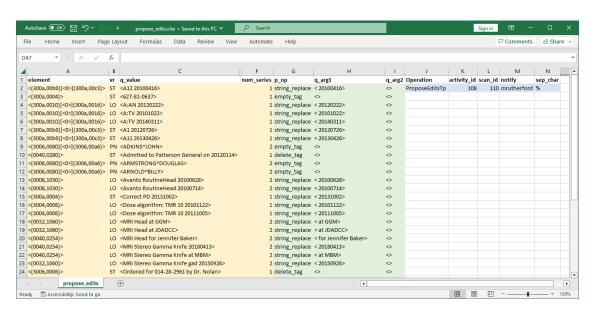
Edit template:

- Blue is supplied by PHI script
- Yellow is cut/pasted from report
- Green is supplied by Curation Mgr.
- Uploaded, ran, and changed confirmed



Scan report:

- Very large (thousands to >1M rows)
- Easily sortable
- Easily scanned by curators
- · Suspicious items highlighted







Future Posda Development

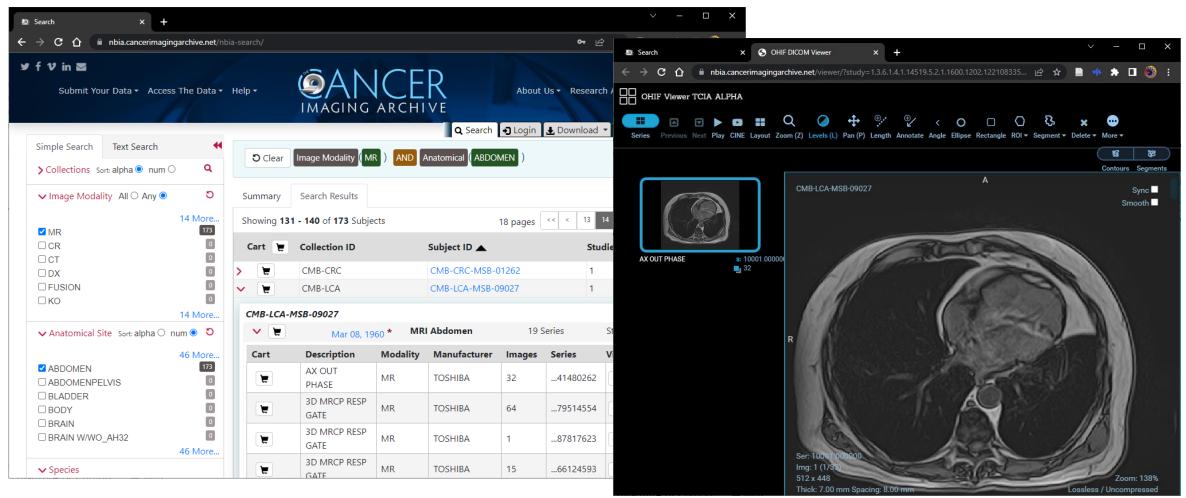
- Improved File Format Support (Both radiology and pathology)
- Enhanced **Defacing** Tools
- Improved Pathology Curation Tools
- Enhanced Visualization Tools
- Improved Clinical Data Curation with Semantic Integration
- Improved Object Store Support
- Improved HPC Support







NBIA (National Biomedical Imaging Archive)







PathDB / Eaglescope / caMicroscope

