

How to make Biomedical Imaging Datasets Al-ready?

Stefan Dvoretskii¹,⁵ // Lucas Kulla¹,⁵ // Philipp Schader ¹,²,⁵ // Josh Moore ⁶ // Marco Nolden ¹,³,⁵

¹ Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg, Division of Medical Image Computing, Germany; ² Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany; ³ Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg, Germany; ⁴ Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany; ⁵ Helmholtz Metadata Collaboration (HMC) Hub Health, Germany; ⁶ Helmholtz Metadata Collaboration (HMC) Hub Health, Germany; ⁶ German

Background

- Al of the new generation (Foundation Models) requires a huge amount of observations
- We need to combine multiple repositories data to get these numbers and prevent bias
- In this setting, manual access to the repositories is not scalable anymore
- We look into the potential improvements for AI-readiness

Public data sources lack conventions for smooth harmonization

Imaging Data Resource

Collecting data on the AI 3.0, "Foundation Model" scale means combining a lot of repositories simultaneously. Repositories in different Bio-/Medical Imaging fields as of now have different structure and access mechanisms. Harmonizing a large dataset variety can be tricky – and involve a lot amendments in the automatic script.



idr.openmicroscopy.org/study Added by: Public data	/info
Sample Type	cell
Organism	Homo sapiens
Study Type	microscopy assay
Imaging Method	confocal microscopy
Imaging Method	fluorescence correlation spectroscopy
Publication Title	A quantitative map of human Condensins provides new insights into mitotic chromosome architecture.
Publication Authors	Walther N, Hossain MJ, Politi AZ, Koch B, Kueblbeck M, Ødegård-Fougner Ø, Lampe M, Ellenberg J
PubMed ID	29632028 https://www.ncbi.nlm.nih.gov/p ubmed/29632028
PMC ID	PMC6028534 https://www.ncbi.nlm.nih.gov/p mc/articles/PMC6028534
Publication DOI	10.1083/jcb.201801048 https://doi.org/10.1083/jcb.201 801048
Release Date	2019-04-15
License	CC BY 4.0 https://creativecommons.org/li censes/by/4.0/
Copyright	Walther et al.
Data Publisher	University of Dundee
Data DOI	10.17867/10000123a https://doi.org/10.17867/10000 123a
Annotation File	idr0052-experimentA- annotation.csv https://github.com/IDR/idr0052 -walther- condensinmap/blob/HEAD/exp erimentA/idr0052- experimentA-annotation.csv



Burden of non machine-actionable sources How to get (meta)data Al-ready?

Open Radiology datasets list snippet.

From missing APIs to "silent" data owners, each pitfall translates in much manual work to obtain the dataset. FAIR data principles can guide through the main caveats.



If a dataset is not machine-actionable, it may be not worth the effort to fetch it at all. If one can not combine it easily with other datasets, it risks staying outside the data pool for the modern AI projects – as it is too much work to invent new ways to harmonize one dataset in a pool of hundreds.

sumset manie (might be antisigoous) [4]	nimai Unique volumes 3D Uni	ue Patients Unique :	D Volumes (N) [4] Anatom	ical Region [5] Main dataset topic	Targets at pixel level [6]	Targets at instance level [7]	largets at image level	Any targets not densely annotated?	Label types (sanity check)	Pathologies annotated?	Non-Pathological ROI annotated?	Provision Status [10]	License [11]	Useable for Scientific	Weights Publishable Data Owner (Email)	Platform [12]	Access policy	Imaging Modalities	Sequences (Separate by ;)	All modalities available for all cases?	World Region (of cente	er) Country (of center)	City (of center)	Scanner (Separate by ;)
DC.	1.010	1.010	1.010 Juna	Lung notule detection		Nodulas	malianant/hanian		PixeLlaval	v	N	Found online (directly downloadable)	CC BY 30			TCIA	Public (downloadable)	CT	CT	*	America (North)	115		
North American Prodrome Longitudinal Study (NAPLS)	1,010	1,010	1,0 10 Lung	Lung nodule detection		Nodues	maignanvoenign		FixeDevel		IN	Found draine (directly downloadable)	00 81 3.0			TCIA	(downoadable)	CT.		1	America (North)	03		
	994	994														IDA Loni	Public							
ACRIN 6667	10,184	969	10,184 Breast	Breast MRI								Found online (directly downloadable)	CC BY 4.0			TCIA	(downloadable)	MR			America (North)	US		
DBIA	4,136	937	4,136 Full Bod	Chinese Cancer Imagi	ng							Found online (access request required)				Own Website	On Request	CT+MR			Asia (South East)	China	Beijing, Guangzhou, Harpin, Shanghai, Tianmen	n
Duke-Breast-Cancer-MRI	5,161	932	5,161 Breast	Breast Cancer	breast				Pixel-level									CT+MR						
AOMIC_ID1000	928	928	Brain									Downloaded	CC BY-SA 4.0	Yes	Yes	OpenNeuro		MR	T1;DWI;fMRI					
Duka Broast Canaar MPI	5 18 1	022	5 181 Proof	buorivo broast concor		Bounding boxes of primary			Bauadian bay	v	v	Equad paling (directly downloadable)	CC BY NC 4 0			TCIA	Public (downloadable)	MD	Non-fatest Tilu: Estest Tilu: 2.4 Part contest	~	Amorico (Morth)	116	Durbon Morth Comline	GE MEDICAL SYSTEMS; MPTronic software;
Longitudinal Evaluation of Familial Frontotemporal Dementia	3, 101	822	a, for breast	invasive breast cancer		lesion			bounding-box			Pound drinne (directly downloadable)	0081-1004.0			TCIA	(downloadable)	MIX	Normal Sal Thir, Patsal Thir, 3-4 Post-contrast		America (North)	03	burnani, North Carolina	JIEMENS
	907	907									-					ICA Loni	_							
AddRET 2022 Challana	1.014	000	1014 518 544	Melanoma, lymphoma	lung Melanoma, lymphoma, lung cancer				Divellaced	v		Encoderation (assessment associated)	TO IA Destricted			TOIA	On Remark	PETACT	BET CT		Europe Allerth	Comen	Tikingan	
Autor ET 2022 Chailenge	1,014	900	1,014 Pt# 800	cancer segmentation	segmentation				PixeHevel	1		Pound online (access request required)	TO IA Restricted			TCIA	On Request	PEI+GI	PELCI	1	Europe (west)	Germany	Tubingen	SIEMENS Biograph mC1
_una 16	888	888	888 Luna	Lung nodule detection		Nodules			Bounding-box	N	N	NotDefined						СТ	ст	Y				
nternational Consortium for Brain Mapping (ICBM)	050	052		11-320												In the second seco								
Max Z; not sure if interesting; slice labels) CT	803	803														ILPS COTI	Public	net en t						
COLONOGRAPHY ACRIN 6664	825	825	Abdomir	al Polyps in colon				Slice indices for polyp locations	Weak (e.g. scribbles)			Found online (directly downloadable)	CC BY 3.0			TCIA	(downloadable)	СТ			America (North)	USA		
SPY2 Breast Dynamic Contrast Enhanced MRI Trial	2,688	719	2,688 Breast	Drug response in brest	t cancer Tumor				Pixel-level			Found online (directly downloadable)	CC BY 4.0			TCIA	(downloadable)	MR	T1;T2					Siemens
Pediatric Brain Tumors Uni HD	700	700	Brain	Tumor segmentation	edema; ce tumor				Pixel-level	Y	N	NotDefined	Custom DUA			Clinical Cooperation	Private	MR	T1; T1ce; T2; FLAIR	Partially	Europe (West)	Germany	Heidelberg	
FastMRI Knee	10,000	700	10,000 Legs	faster.	104				External Labels			Downloaded	Custom DUA			Own Website	On Request	MR	T1, T1pc, T2, FLAIR	N	America (North)	US	NY	
Hecktor MICCAI Challenge	698	698	698 Head an	d Neck Tumor segmentation a outcome prediction	nd				Pixel-level	Y	N	Downloaded & preprocessed/curated	Custom DUA			Grand Challenge	On Request	PET+CT				Canda, France, USA, Switzerland		
·····																								
BraTS2020	2,660	665	2,660 Brain	Brain Tumor Segmenta	ation whole tumor, tumor core, enhancing tum	ior			Pixel-level	Y	N	Downloaded & preprocessed/curated	Custom DUA	Yes	Yes	Own Website	After Registration	MR	T1;T2;T1ce;FLAIR	Y				
DASIS4	663	663	Brain											Yes	Yes	Own Website	Date							
CT Images in COVID-19	661	661	Lung	Covid 19					No Label			Found online (directly downloadable)				TCIA	(downloadable)	СТ						
Cam-CAN Dataset	653	653	Brain	Healthy Brain					No Label			Downloaded	Custom DUA			Own Website	On Request	MR	T1, T2, DWI, MTI, (3x) fMRI, (3x) MEG	N (almost)	Europe (West)	ик	Cambridge	
Advanced-MRI-Breast-Lesions	6,811	632	6,811 Breast	Breast Lesions					Pixel-level + Image-level			Found online (directly downloadable)	CC BY 4.0			TCIA	(downloadable)	MR	T1;T2		America (North)	USA		
Nico) UPENN-GBM	3.301	630	3.301 Brain	Glioblastoma					Pixel-level	Y	Y	Found online (directly downloadable)	CC BY-NC 4.0			TCIA	Public (downloadable)	MR	T1:T2	Y	America (North)	USA	Pennsvivania	
				head and neck squame	ous cell		an the second										Public	07.410			A	1001		
TCIA/HNSCC	627	627	Head an	d Neck carcinoma (HNSCC)	spleen, right kidney, left kidney, gallblad	der.	Clinical data		Image-level			Found online (directly downloadable)	TCIA Restricted			TCIA	(downloadable)	CT+MR			America (North)	USA	MD Anderson	
					esophagus, liver, stomach, aorta, inferio vena cava, pancreas, right adrenal glan	r d.																		
AMOS 2022	800	800	600 Abdomi	Abdominal multi-organ	left adrenal gland, duodenum, bladder,				PixeLlavel	N	N	Drawnloaded	CC BY NC-SA	Vec	Ver	Grand Challanza	After Pagistration	CT+MR		N	Asia (Central & South)	Chipa	Shenthen	
100 2022		000	do Abdomi	segmentation	prostatentents				Tixenevel			Domitobaled	CO DI-NOUR	165	10	Grand Gralenge	Public	CT MILL			Asia (Gentiar & South)	Gina		Philips 3T system , Philips 1.5T system , G
XI Dataset	600	600	Brain	Healthy Brain					No Label			Downloaded	CC BY-SA 3.0			Own Website	(downloadable)	MR	T1, T2, PD, MRA, DTI	Y	Europe (West)	UK	London	1.5T system
				Chronic Obstructive Pu	Imonary																			Somatom Definition AS 40/64/Flash 128; GE Lightspeed VCT 64/ GE Optima 64; Philips
COSYCONET	2,195	563	2,195 Chest	Disease (COPD)			COPD/no-COPD		Image-level	N		Downloaded & preprocessed/curated				Clinical Cooperation	Private	CT+MR	DCE-MRI; inspiratory CT; expiratory CT	Partially	Europe (North)	Germany		Brilliance 64/ iCT 256
																								GE Healthcare Discovery CT750HD; GE Medical
ymph node quantification LNQ23	513	513	Chest	Lymphnode segmentat	ion Lymphnodes	NA	NA	Taining cases only partially annotated	Pixel-level	Y	N	Downloaded & preprocessed/curated	Custom DUA			Grand Challenge	Public (downloadable)	СТ		¥	America (North)	us	Multi-Site (mass general brigham hospitals)	System BrightSpeed; Siemens SOMATOM Definition; Toshiba Aquilion; Philips iCT
EA1141 Abbreviated Breast MRI and Digital Tomosynthesis Mammography in Screening Women With Dense Breasts	1978	500	1976 Breast	Breast cancer detectio	n				No l abel			Found online (directly downloadable)	CC BY 4.0			TCIA	Public (downloadable)	MR			America (North)	USA	Multi-center	
			. Here as carry	Kidney and kidney tum	or												Public				, the local design of the			
KITS2023 Effects of TBL& PTSD on Alzheimer's Disease in Vietnam Vets	489	489	489 Abdomi	al segmentation	kidneys, kidney tumors, cysts				Pixel·level	Y	Y	only preprocessed/curated	CC BY-NC-SA 4.0	Yes	Yes	Own Website	(downloadable)	СТ						
ADNIDOD)	463	463														IDALoni								
ToothFairy	443	443	443 Head an	d Neck CBCT for IAN segment	tation Inferior Alveolar Nerve/Canal				PixeHevel	N	Y	Downloaded & preprocessed/curated	CC BY-SA 4.0			Own Website	After Registration	ст	CBCT	Y	Europe (South)	Italy	Modena	
Generation Study 1 (GS1)	435	435														(BALoni								
NSCLC-Radiomics	422	422	422 Chest	Lung cancer	Gross tumor volume					Y	N	Found online (directly downloadable)	CC BY-NC 3.0			TCIA	Public (downloadable)	СТ	ст		Europe (West)	Netherlands	Maastricht	
CrossModa2022	420	420	420 Brain	Vestibular Schwannom	a Tumor+cochlea		Koos grading		Pixel-level + Image-level	γ	Y	Found online (registration required)				Grand Challenge	After Registration	MR	T1; T2	N	Europe (West)	UK; Netherlands	London; Tilburg	
DASIS1 Huntinaton's Disease NeuroImaging Initiative (HDNI)	416	4 16	Brain											Yes	Yes	Own Website	-							
2 g man (1 m m)	396	396														IDALoni								
Paingen_placebo	395	395	Brain	Brain; Pain fmri								Downloaded		Yes	Yes	OpenNeuro		MR	T1					
				vestibular schwannom	a; nain Brain Matacasis:wastikular sakwana ana											1								
BrainMets GammaKnife-Hippocampal	390	390	Brain	metastasis; Hippocam	pus trigeminal neuralgia	•c			Pixel-level + Image-level	N	Y	NotDefined	TC IA Restricted			TCIA	On Request	CT+MR	T1 ce		America (North)		Multi-Site	Siemens Sonata
Prevent-AD	349	349	Brain													Own Website								
Nournimoning in Ereptatomogral Demontia																								
Neuroimaging in Frontotemporal Dementia	346	346														IDA1 onl								





Tools like triple stores, FAIR Digital Object or a FAIR Data Point come to our rescue here. However, not only the tooling – cultural change is important to convince data owners of publishing their data in a machine-actionable way.



TCIA Portal in Radiology or IDR in Microscopy imaging could serve as role model of the resources with AI-ready data in the Bio/Medical Imaging field. Modern data concepts and universal scheme definition make resources more "fit" for usage with the new generation AI, reducing turnover time and enhancing the model quality.



GERMAN CANCER RESEARCH CENTER

 $\overset{\circ}{\longleftrightarrow} \overset{\circ}{\longleftrightarrow} \overset{\diamond}{\longleftrightarrow} \overset{\diamond}{\bullet} \overset{\bullet$

HELMHOLTZ