Insights from Acquiring Open Medical Imaging Datasets for Foundation Model Development

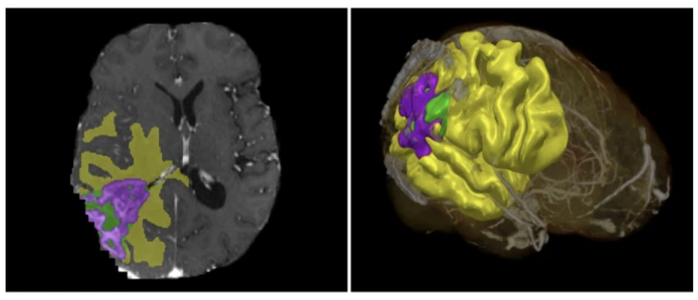
Stefan Dvoretskii [1,5] // Paul Jäger [1,4] // Fabian Isensee [1,4] // Tassilo Wald [1,4] // Constantin Ulrich [1] // Lucas Kulla [1,5] // Philipp Schader [1,2,5] // Klaus Maier-Hein [1,3] // Josh Moore [6] // Marco Nolden [1,3,5]

[1] Division of Medical Computing, Deutsches Krebsforschungszentrum (DKFZ) Heidelberg German Cancer Research Center (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 University Hospital, Heidelberg, Germany; [4] Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany; [5] Helmholtz Metadata Collaboration (HMC) Hub Health, German Cancer Research Center (DKFZ), Heidelberg, Germany; [6] German BioImaging, University of Konstanz, Germany

Background

- Foundation Model for Radiology involves a huge amount of clinical imaging volumes
- We made a list of open clinical imaging datasets – UK Biobank, NAKO, HCP and others
- Some open clinical imaging datasets were not machine-actionable and accessible
- We look into the problems and suggest helpful concepts

Data for Radiological Foundation Model



MedSAM CT image datasets (Isensee, Jäger et al. Nat Methods 2021)

Dataset Name	Modality	Segmentation Targets	# of scans
AbdomenCT-1K [1], [2]	CT	Liver, kidneys, pancreas, spleen	1056
Adrenal-ACC-Ki67* [3]–[5]	CT	Adrenocortical carcinoma	53
AMOS-CT [6]	CT	Abdominal organ	200
AutoPET [7]	PET-CT	Whole-body tumor	900
COVID-19 Seg. Challenge [8], [9]	CT	COVID-19 infections	199
COVID-19-CT-Seg [10]	CT	COVID-19 infections, left lung, and right lung	20
GLIS-RT [11]	CT	Head tumor	75
HCC-TACE-Seg* [5], [12]	CT	Liver cancer	70
HECKTOR [13]	PET-CT	Head and neck tumor	524
INSTANCE [14]	CT	Hematoma	100
KiPA [15], [16]	CT	Kidney, tumor, renal artery, renal vein	70
KiTS [17]	CT	Kidney, tumor, cyst	489
LNQ2023 * [18]	CT	Mediastinal lymph node	393
Lymph Nodes [19], [20]	CT	Lymph nodes	176
MSD-Colon Tumor [21]	CT	Colon tumor	126
MSD-Hepatic Tumor [21]	CT	Hepatic tumor	303
MSD-Lung Tumor [21]	CT	Lung tumor	96
MSD-Pancreas [21]	CT	Pancreas, pancreas tumor	281
MSD-Spleen [21]	CT	Spleen	61
NSCLC Pleural Effusion [5], [22], [23]	CT	Pleural effusion	78
NSCLC Radiogenomics [24]	CT	Lung Tumor	88
ORG [25]	CT	Whole-body organs	140
SegTHOR [26]	CT	Esophagus, heart, aorta, trachea	40
StructSeg* [27]	CT	Nasopharyngeal cancer and lung cancer, with OAR and GTV	50
TotalSegmentator [28]	CT	Whole body organs	1204
WORD* [29]	CT	Abdominal organs	150

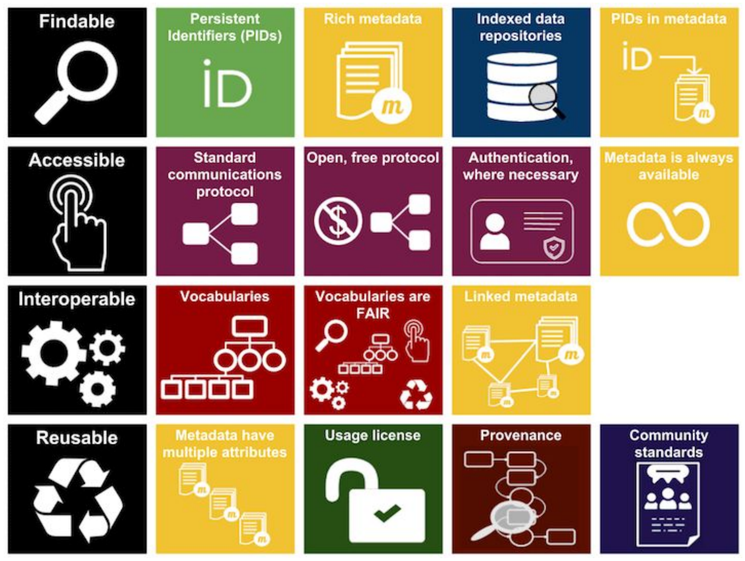
MedSAM CT image datasets (Ma et al. Nat. Communications, 2024)

Our data list contained observations count, imaging targets, anatomical regions, pathology, provisioning and license information; data owner contact, imaging modalities information, number of sequences, geographical region and scanner metadata, as well as free-format remarks.

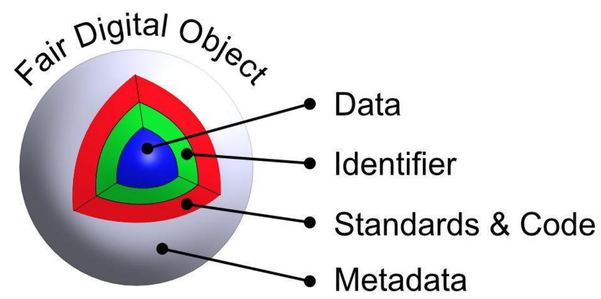
Datasets are not machine-actionable

We ended up doing a lot by hand. This was time and effort consuming.

A lot of described problems could be seen through the **FAIR** data guidelines lens.



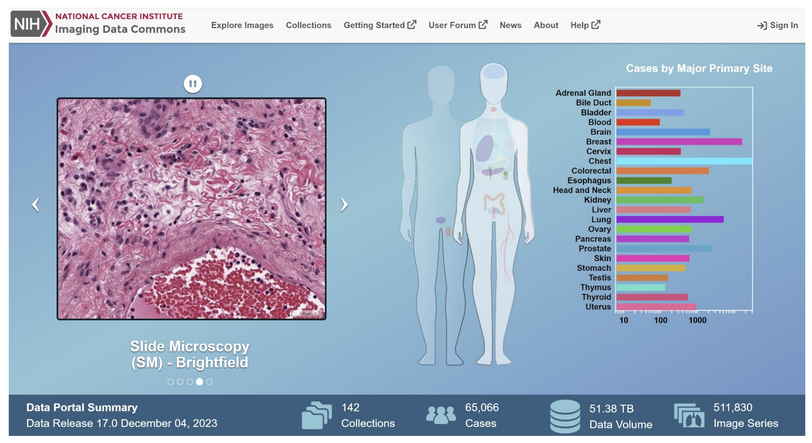
One of the FAIR guidelines summaries. Source: Australian Research Data Commons

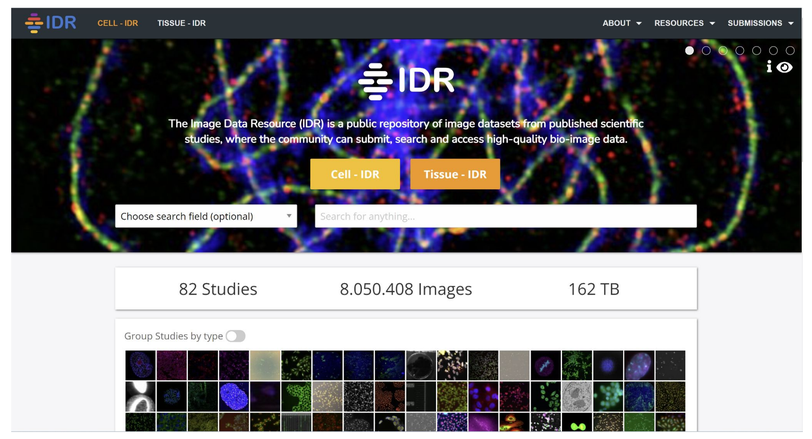


FAIR Digital Object schema

How to get (meta)data AI-ready?

Data concepts like **FAIR Digital Object** or a **FAIR Data Point** provide basis for machine data access. Some data portals provide good data overview mechanisms. There may even be FAIR-compliant APIs behind.





Foundation Models require a lot of data. It is instructive that public data is machine-actionable. With modern developments like **data spaces** and **trusted research environments**, we should define what AI-ready means and how to organize data curation for all stakeholders.

	Dataset Name (might be ambiguous)	umes (N)	Anatomical Region	Main dataset topic	Targets at pixel level	Targets at instance level	Targets at image level	Any targets not densely annotated? (e.g. only a subset lesions)	Label types (sanity check)	Pathologies annotated?	Non Pathological ROI annotated?	Provision Status	License	Useable for Scientific publication	Weights Publishable	Data Owner (Email)	Platform	Access
41	LIDC	1,010	Lung	Lung nodule detection		Nodules	malignant/benign		Pixel-level	Y	N	Found online (directly downloadable)	CC BY 3.0				TCIA	Public (downloa
42	North American Prodrome Longitudinal Study (NAPLS)																IDA Loni	
43	ACRIN 6667	10,104	Breast	Breast MRI								Found online (directly downloadable)	CC BY 4.0				TCIA	Public (downloa
44	OBIA	4,156	Full Body	Chinese Cancer Imaging								Found online (access request required)					Own Website	On Requ
45	Duke Breast Cancer MRI	5,161	Breast	Breast Cancer	breast				Pixel-level									
46	AOMIC ID1000		Brain									Downloadec	CC BY-SA 4.0	Yes	Yes		Open Neuro	
47	Duke-Breast-Cancer-MRI	5,161	Breast	Invasive breast cancer		Bounding boxes of primary lesion			Bounding-box	Y	Y	Found online (directly downloadable)	CC BY-NC 4.0				TCIA	Public (downloa
48	Longitudinal Evaluation of Familial Frontotemporal Dementia																IDA Loni	
49	AutoPET 2022 Challenge	1,014	Full Body	Melanoma, lymphoma, lung cancer segmentation	Melanoma, lymphoma, lung cancer segmentation				Pixel-level	Y		Found online (access request required)	TCIA Restricted				TCIA	On Requ
50	Luna16	888	Lung	Lung nodule detection		Nodules			Bounding-box	N	N	NotDefined						
51	International Consortium for Brain Mapping (ICBM)																IDA Loni	
52	(Max Z, not sure if interesting, slice labels) CT COLONOGRAPHY | ACRIN 6664		Abdominal	Polyps in colon				Slice indices for polyp locations	Weak (e.g. scribbles)			Found online (directly downloadable)	CC BY 3.0				TCIA	Public (downloa
53	ISPY2 | Breast Dynamic Contrast Enhanced MRI Trial	2,688	Breast	Drug response in breast cancer	Tumor				Pixel-level			Found online (directly downloadable)	CC BY 4.0				TCIA	Public (downloa
54	Pediatric Brain Tumors Uni I D		Brain	Tumor segmentation	edema ce tumor				Pixel-level	Y	N	NotDefined	Custom DUA				Clinical Cooperation	Private
55	FastMRI Knee	10,000	Legs	make MRI scans up to 10X faster					External Labels			Downloadec	Custom DUA				Own Website	On Requ
56	Hecktor MICCAI Challenge	698	Head and Neck	Tumor segmentation and outcome prediction					Pixel-level	Y	N	Downloadec & preprocessed/curated	Custom DUA				Grand Challenge	On Requ
57	BraTS2020	2,660	Brain	Brain Tumor Segmentation	whole tumor, tumor core, enhancing tumor				Pixel-level	Y	N	Downloadec & preprocessed/curated	Custom DUA	Yes	Yes		Own Website	After Registrat
58	OASIS4		Brain											Yes	Yes		Own Website	
59	CT Images in COVID-19		Lung	Covid-19					No Label			Found online (directly downloadable)					TCIA	Public (downloa
60	Cam-CAN Dataset		Brain	Healthy Brain					No Label			Downloadec	Custom DUA				Own Website	On Requ
61	Advanced-MRI-Breast-Lesions	6,811	Breast	breast Lesions					Pixel-level + Image-level			Found online (directly downloadable)	CC BY 4.0				TCIA	Public (downloa
62	(Nico) UPENN GBM	3,301	Brain	Glioblastoma					Pixel-level	Y	Y	Found online (directly downloadable)	CC BY-NC 4.0				TCIA	Public (downloa
63	TCIA-HNSCC		Head and Neck	head and neck squamous cell carcinoma (HNSCC)			Clinical data		Image-level			Found online (directly downloadable)	TCIA Restricted				TCIA	Public (downloa
64	AMOS 2022	600	Abdominal	Abdominal multi-organ segmentation	spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus				Pixel-level	N	N	Downloadec	CC BY-NC-SA	Yes	Yes		Grand Challenge	After Registrat
65	IXI Dataset		Brain	Healthy Brain					No Label			Downloadec	CC BY-SA 3.0				Own Website	Public (downloa

Our dataset list snippet. Some columns and rows, like scanner and geographical information, are not shown.