HELMHOLTZ Metadata Collaboration

MIC MEDICAL IMAGE COMPUTING

HELMHOLTZ IMAGING

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION
Research for a Life without Cancer

AST | E&E | ENERGY | **HEALTH** | INFORMATION | MATTER | DATA COMMONS | HMC OFFICE | HMC PROJECTS

Insights from Acquiring Open Medical Imaging Datasets for Foundation Model Development

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

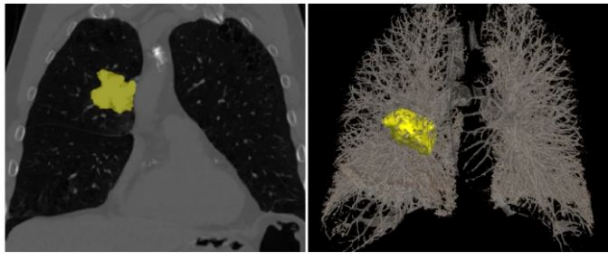
[1] DKFZ Division of Medical Computing; [2] DKFZ Interactive Machine Learning; [3] Pattern Analysis and Learning Group, Heidelberg University Hospital; [4] Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany; [5] Helmholtz Metadata Collaboration (HMC) Hub Health, DKFZ, Heidelberg, Germany; [6] German BioImaging e.V., University of Konstanz, Germany

Helmholtz Munich

www.helmholtz-metadaten.de

Foundation Model for Radiology

- Aid radiological research and clinical practice
- Variety of tasks in the domain
- Example: Segmentation of pathologies
- Clinical: Early detection, Therapy response monitoring
- Problem: “real” clinical imaging data is difficult to share! (in Germany)



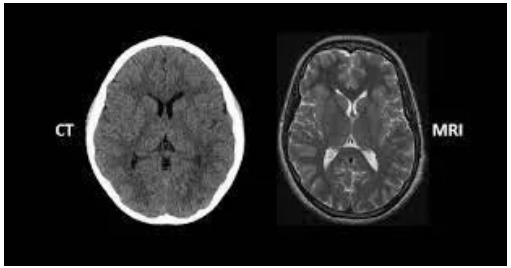
Example task: lung nodules segmentation

Source: Isensee, Jäger et al. Nat Methods 2021

RACOON
Die Radiologie Kooperation im NUM

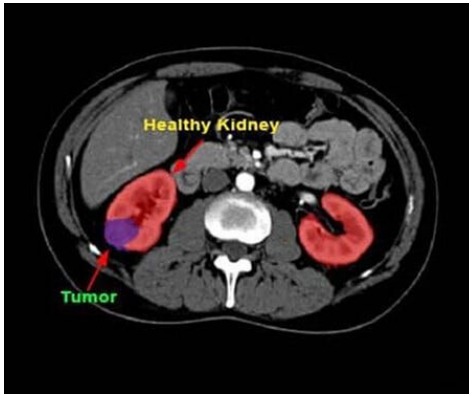
We were looking for diverse and big open clinical imaging datasets

- Different body parts
- Different modalities (CT, MR)
- Different institutions and geographic regions
- Validation subset: high-quality annotations



Source: [1]

Source: [2]

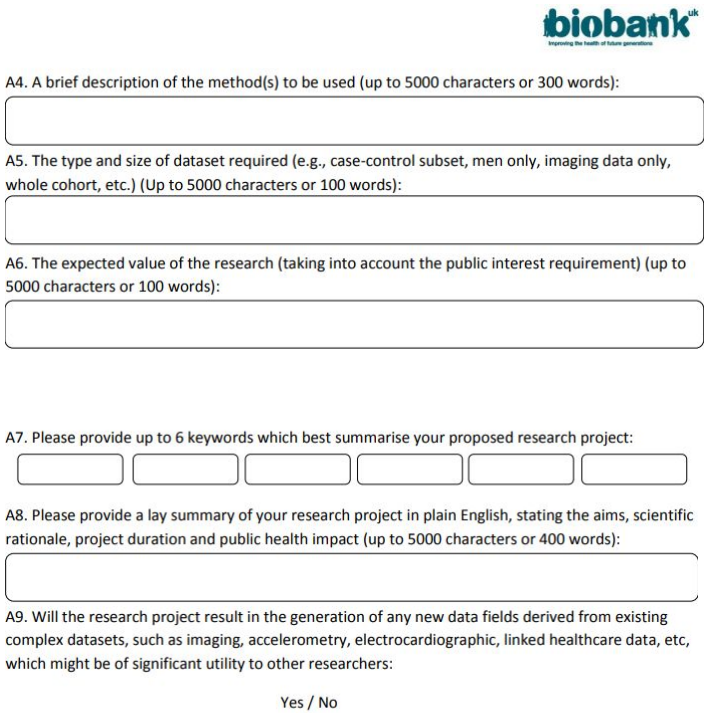


Source: [3]

1. https://www.maximedturkey.com/de/blog/ct-scan-vs-mri-what-s-the-difference
2. https://www.freepik.com/photos/world
3. https://medium.com/vsinghbisen/what-is-medical-image-annotation-role-in-ai-medical-diagnostics-a44338bb9bdb

We ended up doing a lot by hand

- Fetching metadata
- Portal registrations
- Access requests
- License evaluation



biobank uk

A4. A brief description of the method(s) to be used (up to 5000 characters or 300 words):

A5. The type and size of dataset required (e.g., case-control subset, men only, imaging data only, whole cohort, etc.) (Up to 5000 characters or 100 words):

A6. The expected value of the research (taking into account the public interest requirement) (up to 5000 characters or 100 words):

A7. Please provide up to 6 keywords which best summarise your proposed research project:

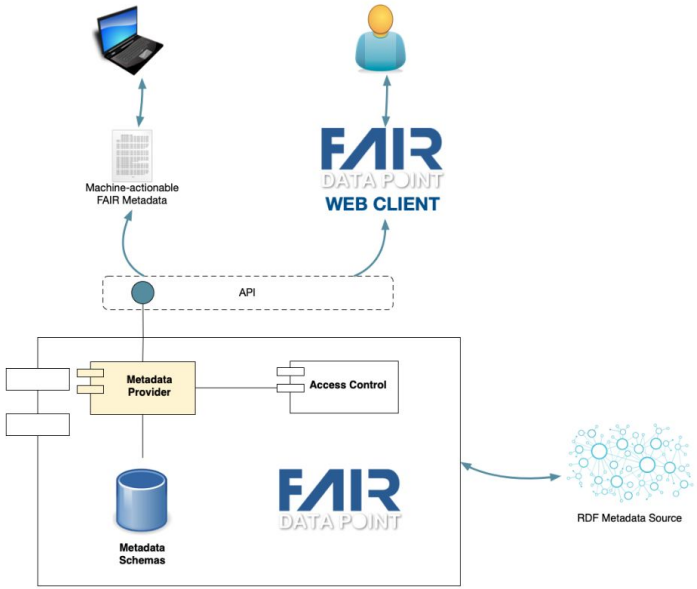
A8. Please provide a lay summary of your research project in plain English, stating the aims, scientific rationale, project duration and public health impact (up to 5000 characters or 400 words):

A9. Will the research project result in the generation of any new data fields derived from existing complex datasets, such as imaging, accelerometry, electrocardiographic, linked healthcare data, etc, which might be of significant utility to other researchers:

Yes / No

How to make data fetching more machine-actionable?

- Efficient and transparent access requests
- Efficient exposure of data and metadata
- FAIR Data Point, FAIR Digital Object
- FAIR *should be* machine-actionable



High-level architecture of a FAIR Data Point

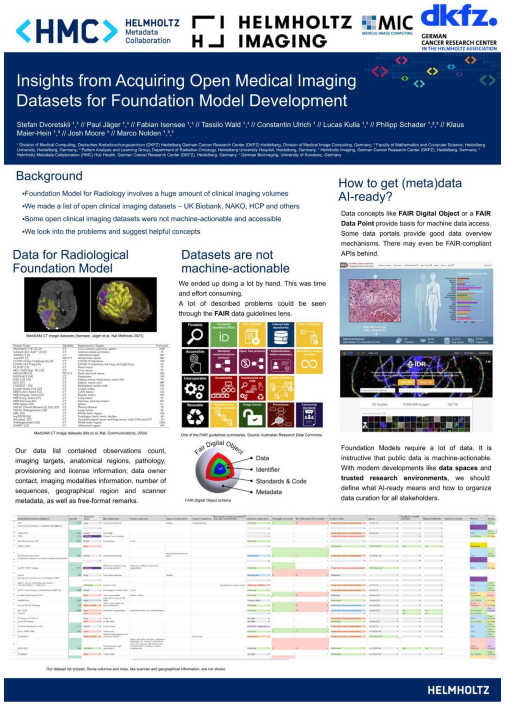
https://specs.fairdatapoint.org/fdp-specs-v1.2.html

Modern initiatives transform public data landscape

- Data spaces
- Trusted research / Secure processing environments
- The way data is handled evolves
- Metadata is getting more important

Conclusion and open questions

- Open Medical Imaging data is rich
- Preparing it for Foundation Model development poses machine-actionability problems
- How to assure open imaging data is AI-ready?
- What are the relevant metadata for Foundation Models?



Thank you for your attention