How to make Biomedical Imaging Datasets AI-ready?

Stefan Dvoretskii[1,5] // Lucas Kulla[1,5] // Philipp Schader [1,2,5] // 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

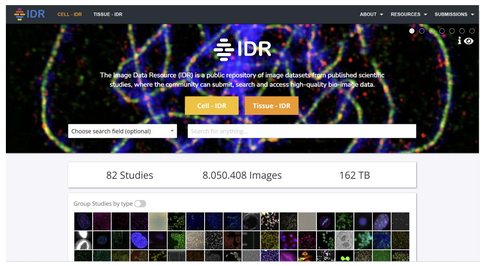
Background

- AI 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

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.

Imaging Data Resource study annotation



idr.openmicroscopy.org/study/info **Added by: Public data**	
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/pubmed/29632028
PMC ID	PMC6028534 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6028534
Publication DOI	10.1083/jcb.201801048 https://doi.org/10.1083/jcb.201801048
Release Date	2019-04-15
License	CC BY 4.0 https://creativecommons.org/licenses/by/4.0/
Copyright	Walther et al.
Data Publisher	University of Dundee
Data DOI	10.17867/10000123a https://doi.org/10.17867/10000123a
Annotation File	idr0052-experimentA-annotation.csv https://github.com/IDR/idr0052-walther-condensinmap/blob/HEAD/experimentA/idr0052-experimentA-annotation.csv

Open Radiology datasets list snippet.

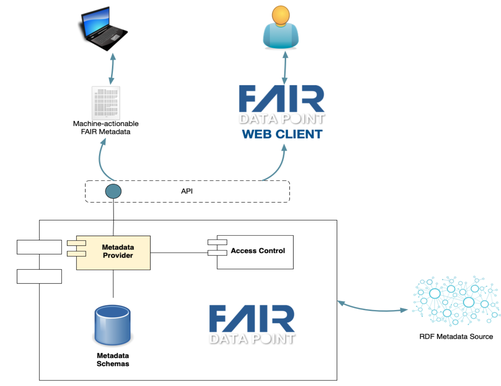
Burden of non machine-actionable sources

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.

How to get (meta)data AI-ready?

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.



FAIR Data Point structure. Source: fairdatapoint.org

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.