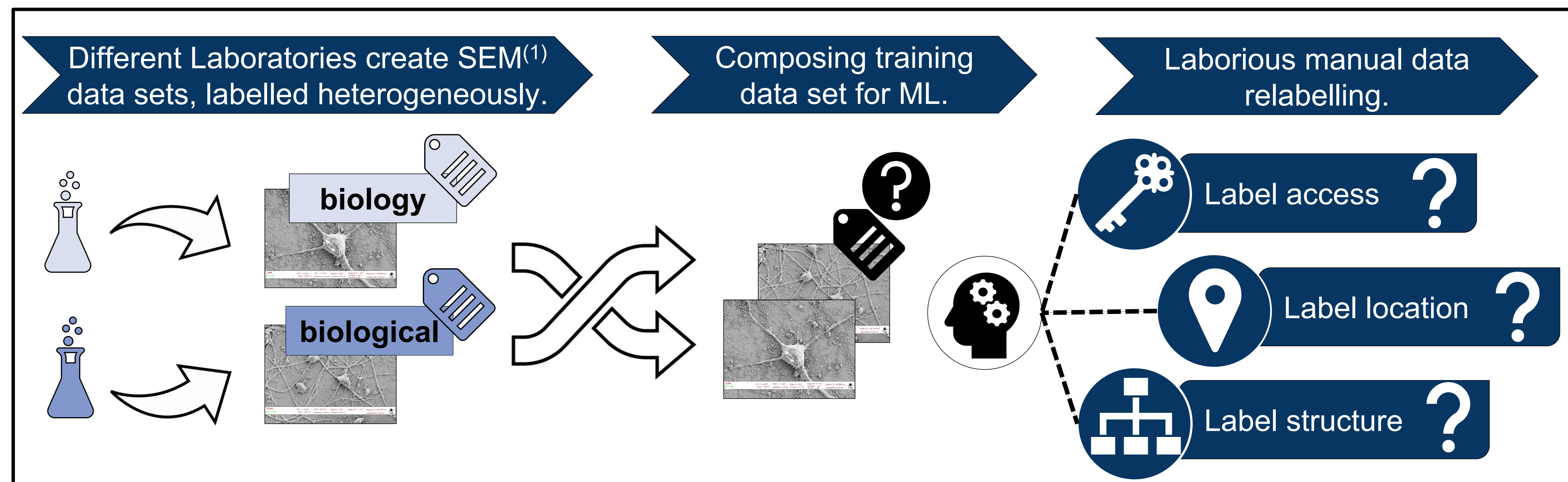# FAIR DO Application Case for Composing Machine Learning Training Data

Nicolas Blumenröhr, Thomas Jejkal, Andreas Pfeil, Rainer Stotzka
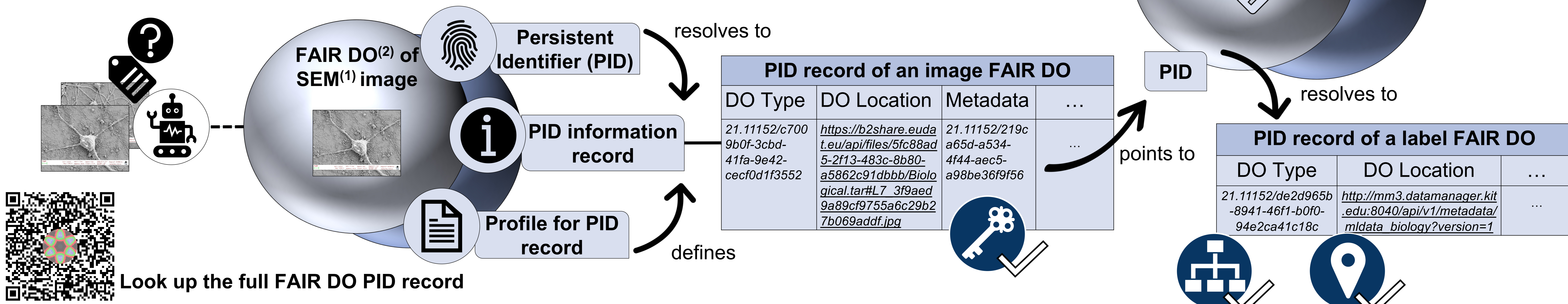
Composing Machine Learning (ML) training data sets from heterogeneous sources is laborious due to their relabelling into uniform categories.

Different Laboratories create SEM[1] data sets, labelled heterogeneously.

biology

biological

Composing training data set for ML.

Laborious manual data relabelling.

Label access ?

Label location ?

Label structure ?

Conclusions
- Images and their labels represented as FAIR DOs can be located and accessed easier.
- FAIR DOs enable machine actionable decisions on the data.
- This supports automated relabelling and saves a lot of time for the scientist.

To automate this task, the FAIR Digital Object (FAIR DO) concept can be used.

Composing ML training data from SEM[1] data sets represented as FAIR DOs.

Data relabelling using PID- and type-based computer actionability for FAIR DOs.

FAIR DO[2] of label metadata

FAIR DO[2] of SEM[1] image

Persistent Identifier (PID)   resolves to

PID information record

Profile for PID record   defines

PID

resolves to

PID record of an image FAIR DO			
DO Type	DO Location	Metadata	…
21.11152/c7009b0f-3cbd-41fa-9e42-cecf0d1f3552	https://b2share.eudat.eu/api/files/5fc88ad5-2f13-483c-8b80-a5862c91dbbb/Biological.tar#L7_3f9aed9a89cf9755a6c29b27b069addf.jpg	21.11152/219ca65d-a534-4f44-aec5-a98be36f9f56	…

points to

PID record of a label FAIR DO		
DO Type	DO Location	…
21.11152/de2d965b-8941-46f1-b0f0-94e2ca41c18c	http://mm3.datamanager.kit.edu:8040/api/v1/metadata/mldata_biology?version=1	…

Look up the full FAIR DO PID record

(1) **Scanning Electron Microscopy (SEM)** data set, provided by R. Aversa et. al. http://doi.org/10.23728/b2share.19cc2afd23e34b92b36a1dfd0113a89f
(2) Introduction to PIDs and FAIR DOs: https://kit-data-manager.github.io/fairdo-cookbook/about.html